Developing Toponym Classification and Geolocation Estimation from Thai Tweets

Chalamkate, T.,[1] Tinnachote, C.[1] and Rutherford, A. T.[2*]
[1]Department of Survey Engineering, Faculty of Engineering, Chulalongkorn University, Phayathai Road, Wangmai, Pathumwan, Bangkok 10330, E-mail: tuvachitchalamkate@gmail.com, chanin.ti@chula.ac.th
[2]Department of Linguistics, Faculty of Arts, Chulalongkorn University, Phayathai Road, Wangmai, Pathumwan, Bangkok 10330, E-mail: tuvachitchalamkate@gmail.com, Attapol.T@chula.ac.th
**Corresponding Author*

DOI: https://doi.org/10.52939/ijg.v19i7.2741

Abstract
Twitter is a highly active microblogging platform globally, and in Thailand, it has secured the 10th rank for the highest user base in 2021. The platform is known for its rapid news dissemination capabilities and richness in topological and geolocation information. As such, the extraction of geospatial data from unstructured text has become an emerging field that has caught the attention of geospatial researchers. Geospatial extraction aims to leverage the location-based information present in textual data, which can be achieved through two methods, including toponym extraction and geocoding. The first involves identifying the geographic name from the text, and the second involves assigning the corresponding coordinates to the identified location. This study presents two key contributions. Firstly, a transformer-based tool was developed for extracting geo-names utilizing the BERT architecture, achieving an F1 score of 0.919 for overall accuracy. Secondly, the geocoding task was explored. The primary focus was to estimate the location of extracted geographic names that could not be matched with established online databases such as Google Geocoding, implying that the geographic names may not have existed or might not have been accurately identified. To achieve this objective, a geographic name dataset sourced from Twitter was compiled and utilized as a test dataset. The proposed approach involved the application of a clustering machine learning model, along with the utilization of topological properties, to develop a model for estimating the location of geographic names. The efficiency of the proposed method was assessed using Root Mean Square Error (RMSE) to compare the geographic coordinates of the estimated location from the model with the actual geographic coordinates of the location. Results revealed that the topology words model exhibited extremely high efficiency, with an RMSE of 0.947 km.

Keywords: BERT**,** Geocoding**,** Geospatial, Toponym extraction, Topology words

1. Introduction

Location-based information can be extracted in textual form from diverse sources, such as social media, online news websites, housing sale advertisements, and restaurant reviews. These sources are increasingly crucial in geospatial applications as they provide a wide variety of data, including descriptions and narratives of events (e.g., disasters) or location information within the text [1]. Social media is an indispensable communication tool that enables fast, frequent, and extensive information sharing. In Thailand, it serves as the primary news source for 78% of the population, the highest proportion worldwide. Twitter is also a notable social media platform in Thailand, ranking 10th globally in terms of its user base and serving as the second most popular platform for sharing news after Facebook [2]. Hence, there are significant opportunities for diverse applications, given that over 60% of digital data available on social media, weblogs, and other platforms contain geo-references [3]. As previously discussed, the first step in geoparsing is to extract toponyms from text using tools that typically incorporate the Name Entity Recognition (NER) technique. However, studies on NER for the Thai language are limited compared to other popular languages such as English, French, and Chinese.

In the various NER models were developed to extract information such as people, organizations, places, times, and emails. However, it should be noted that these models also extracted toponyms, which are place names.

Unfortunately, the toponyms obtained from these tools were not categorized, making it difficult to identify locations in online databases such as Google geocoding when using the extracted data to collect important name information. The lack of categorization also made it challenging to perform spatial analysis, such as creating a database for navigation systems requiring grouped individual geographic names or estimating toponyms' coordinates. To improve estimation accuracy, some studies recommended using administrative zones to narrow the scope [4] and [5] In response to the need, this study developed a toponym recognition model using WangchanBERTa, specifically for its second aim, geocoding.

To achieve the second aim of this study, geocoding or toponym resolution was employed to reduce the ambiguity of the toponyms extracted from the first step by retrieving their corresponding coordinates from the geographic encyclopedia database, also known as the Gazetteer. Previous studies on geolocation estimation from social media data have identified various spatial indicators, including the seven indicators identified by Zheng et al., [6] (1) referenced locations in messages, (2) social networks, (3) user profiles, (4) geotags, (5) third-party sources, (6) time zone, and (7) IP addresses.

Drawing on the findings of the earlier studies, this study aimed to estimate the geolocations of toponyms mentioned in Twitter messages. These toponyms were extracted using the toponym recognition model developed in the first part of the study. To ensure the quality of the extracted data, predefined rules were employed to filter out anticipated typos or incomplete information that may have arisen from the use of the toponym recognition tool. Subsequently, the filtered data was linked to the Google Geocoding API, an online database known for its comprehensive and accurate information about Thai names [7]. In addition to reducing the ambiguity of toponyms, this study also aimed to estimate the geolocations of places where coordinates could not be determined by utilizing the topology words programming technique. To accomplish this, four commonly used machine learning clustering techniques in spatial analysis, namely DBSCAN, K-means, K-medoids, and Agglomerative clustering, were employed. After comparing and summarizing the results obtained using each technique, the study selected the one that provided the most accurate results. This study was conducted with two aims:

1. To develop a toponym recognition model for Thai Twitter data

2. To develop an algorithm that estimates the geolocations of toponyms that cannot be geocoded according to online databases.

Following this 1) Introduction section, the remainder of the study is organized into sections including 2) Literature Review 3) Process and methodology, 4) Results, 5) Discussion, and 6) Conclusion.

The Toponym Classification and Geolocation Estimation source code, training data, and annotated test data are available at https://github.com/crescendonow/thai geoparsing.

2. Literature Review

2.1 Thai Named Entity Recognition Research

Chanlekha et al., [8] proposed a hybrid approach using statistical methods and a heuristic rule-based model to generate rules based on name entities to improve the accuracy of toponym extraction. The model produced promising results for magazine writing but had a suboptimal performance for newspaper articles, possibly due to the formal and pattern-oriented writing style. In a subsequent attempt to overcome the limitations of toponym extraction in the Thai language, Chanlekha and Kawtrakul [9] combined a heuristic rule-based model, a dictionary of specific terms, and the Maximum Entropy or Logistic Regression model using tokenization.

Their approach was practical for some name entity types, such as personal names, but less effective for organization names due to data fragmentation. Note that the efficiency of calculating function weights was reduced by data fragmentation, and recognizing name entities that spanned two words before or after a reference word resulted in lower accuracy (0.776) than recognizing name entities that were one word before or after a reference word (0.8987). Furthermore, place-related name entities were found to have the least accuracy, likely due to their greater ambiguity when compared to other types of name entities. Later in the timeline, Tirasaroj and Aroonmanakun [10] developed a NER system for the Thai language using the CRF model. The study utilized the "BEST 2009" archive from NECTEC, which comprised approximately 90,000 words. The experiment compared the CRF model training between word and syllable segmentation data.

The experimental results showed that the overall accuracy of the word and syllable segmentation models was similar, with accuracy values of 0.8039 and 0.808, respectively. However, the syllable segmentation model outperformed the word segmentation model for location-related name entity

recognition, with accuracy values of 0.7692 and 0.7372, respectively. Recently, Thattinaphanich and Prom-On [11] constructed a NER system using an inverted neural network called Bi-Directional Long Short Term Memory (Bi-LSTM) with CRF and 13 data layers, and the location layer is one of them. Results revealed that the best-performing model achieved an overall accuracy of 0.8773. The latest and most advanced model for Thai language NER is WangchanBERTa, which results from implementing transfer learning techniques on a Thai language dataset of over 78.5 GB using the bidirectional encoder representations from transformers (BERT) architecture. This model is considered the most efficient and advanced among the previously discussed models [12].

2.2 Textual Geolocation Estimation

Xu et al., [13] proposed the Location Propagation Probability algorithm to predict a user's location using social networking and user location information on Weibo, a Chinese social media platform, which achieved an accuracy of 68.2% at the city level and 73.7% at the provincial level. Additionally, Williams et al., [14] analyzed location data from user addresses and social networks on Twitter using an algorithm that employed density-based clustering algorithm and noise (DBSCAN) in conjunction with K-means to cluster spatial relationships and predict geocoordinates, achieving an accuracy of 30% at the location level within a five-kilometer radius.

3. Study Area and Data

This study employed a rigorous methodology, as illustrated in Figure 1. The process began with creating a corpus of 28,082 messages, which were utilized for training the toponym recognition model. Next, the dataset was split into three parts: the train set, validate set, and test set, with an 80%, 10%, and 10% allocation, respectively. After training and evaluating the performance of the toponym recognition model, it was applied to 100 datasets obtained from Twitter that were not part of the original corpus. The toponyms extracted were then subject to additional rule-based processing and used for geocoding to estimate the corresponding place locations using the topology words algorithm. Eventually, the algorithm's accuracy was compared with the results obtained using four commonly used clustering models, and the estimated locations of the toponyms were visualized on a map.

3.1 Data Collection from Twitter

The geographic focus of this study was on Bangkok metropolitan and its vicinity, specifically including Nakhon Pathom, Nonthaburi, Pathum Thani, Samut Prakan, and Samut Sakhon, which have the highest population density in Thailand. To ensure that only relevant data were obtained, the study utilized a bounding box (BBOX) as the location parameter for the endpoint filter to define the spatial scope. This BBOX consisted of the lower left and top right corner coordinates. Texts were then processed to determine whether they contained coordinate values within the BBOX. In cases where no coordinates were identified, the toponyms data in the place field were used to determine whether the region specified remained within the BBOX territory. Messages that did not contain either coordinates or toponyms were disregarded to ensure that the data collected were accurate and relevant. The data used in this research consists of two main parts: the data used to train the model and the data used as test data to estimate the geolocations of places.

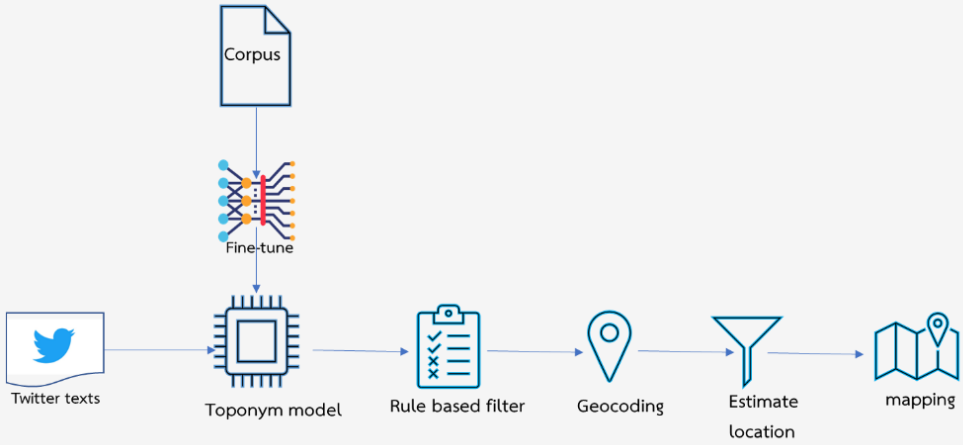

Figure 1: The geoparsing process for Thai Twitter data

3.1.1 Toponym recognition

The toponym recognition model was developed using a dataset of 28,082 Twitter messages containing 1,974,211 characters collected between September and November 2019. The Tweepy command library in Python was utilized to download the data, and a BBOX was established using the latitude and longitude coordinates of Bangkok metropolitan and its vicinity. This BBOX was defined by the bottom left (BL) and top right (TR) geographic coordinates, respectively, using a list of coordinates [LonBL, LatBL, LonTR, LatTR], such as [100.060422, 13.434143, 101.014372, 14.189893]. The data were filtered based on the BBOX, and only tweets within its geographic boundaries were retained. The filtered data were subsequently saved as a JSON file.

3.1.2 Twitter data

In this study, Twitter data were used as test data to estimate the geolocation of a place using Tweepy in conjunction with Google Custom Search API and web scraping techniques. In addition, due to the limited timeframe for accessing data from the free Twitter API, other techniques were utilized to maximize data acquisition. Finally, 100 datasets of location-related information were collected, with each set containing 5-20 messages stored as JSON files.

3.2 Language Corpus Development

3.2.1 Data annotation

In this study, the toponyms were categorized into 18 subcategories under five core categories: 1) natural locations, 2) buildings, 3) administrative zones, 4) locations outside of Thailand, and 5) other locations. The categorization process was carried out using human annotation, which involved creating a guideline to assist in selecting the appropriate annotation for each type of toponym. Table 1 presents examples of the annotation formats used for each category to demonstrate the annotation process.

3.2.2 Sequence tagging using IOB tags

In IOB tag construction, tags are used in a sequence tagging process, where “B” represents the beginning of a noun phrase, “I” represents the inside of the current noun phrase, and “O” represents other words or tokens that are not part of a name entity and are not extracted [15]. Different word segmentations can result in different word subunits. For example, in Text 1 of Table 1, “@Ordinary Champ: The view of <NAT>Mekong River in Nakhon Phanom</NAT> is the most beautiful. I like it very much.”, the segments can be labeled as (@Ordinary_Champ, O), (The view of, O), (River, B-NAT), (Mekong, I-NAT), (Nakhon Phanom, I-NAT), (Beautiful, O), (most, O), (and, O), (I, O), (like it very much, O). In this example, “Mekong River in Nakhon Phanom” represents the toponym that needs to be extracted, and it consists of three subunits: River, Mekong, and Nakhon Phanom, where “River” is the first unit of the toponym in Thai, and “Mekong” and “Nakhon Phanom” are the components of the toponym, respectively. The remaining words are labeled as “O.”

Table 1: Examples of toponym categories and annotation formats

No.	Message
1	@Ordinary_Champ วิว<NAT>แม่น้ำโขงนครพนม</ NAT>สวยสุดละนี่ชอบมากกกกกก
	@Ordinary_Champ: The view of <NAT>Mekong River in Nakhon Phanom</NAT> is the most beautiful. I like it very much.
2	141062 @7.57 ฝนตกหนักเลย <ROAD>แจ้งวัฒนะ</ROAD> น่าจะตกก่อนออกจากบ้านอุตสาห์จะไปเดินออกกำลังกาย<RCT>สวนจตุจักร</RCT> ท่าจะไม่รอด 😂
	141062 @7.57 It’s raining heavily at
	<ROAD>Chaengwattana</ROAD>.It should have rained before I left the house. I’m trying to go for a walk and exercise at the <RCT>Chatuchak Park</RCT> here. It seems I won’t make it. 😂
…	……
3.	ตักบาตรเข้าสารแห้ง... เราก้อทำน่ะ 👍🏻 🤗 🤗 — ที่ <MKT>ตลาดน้ำขวัญ-เรียม</MKT> / <RP>วัดบางเพ็งใต้+วัดบำเพ็ญเหนือ กรุงเทพมหานคร</RP>
	Offering raw white rice to monks... I also do it 👍🏻 🤗 🤗 — at <MKT>Kwan-Riam Floating Market</MKT> / <RP>Wat Bang Peng Tai+Wat Bampen Nuea, Bangkok</RP>

4. Methods

4.1 Constructing the Toponym Recognition Model

The approach employed for constructing the toponym recognition model in this study was adapted from Thai NER research. However, academic knowledge of Thai NER remains scarce compared to more widely studied languages like English, French, and Chinese.

In 2018, Google AI Language introduced BERT, an approach that utilizes an encoder-specific knowledge transfer architecture to develop a language model. This approach involves training the model on unlabeled datasets from Wikipedia and a book corpus comprising published English books. The versatility of BERT allows it to be fine-tuned for a broad range of natural language processing problems, including NER [16]. However, there are some limitations for the Thai language due to the model being trained on data from over 100 languages simultaneously, which makes it challenging to capture the specifics of the Thai language or account for the diverse range of topics that appear in the Thai language datasets. To address this limitation, RoBERTa was developed for the Thai language in 2021 and trained on over 78.5 GB of Thai language datasets from various sources [12].

4.1.1 BERT architecture

BERT architecture for extracting and classifying geo-names as previously discussed, BERT leverages knowledge transfer and attention mechanisms to learn the relationships between words or sub-words in text. Typically, a knowledge transfer system comprises an encoder that receives input text and a decoder that predicts the result. However, in this study, only the encoder component of BERT was utilized, and a classifier element was added instead. The use of BERT in this study is summarized in Figure 2. In Figure 2, the first section after the input sequence is called embeddings, which consists of three subsections: position, segment, and token. As shown in the example, the input sentence was divided into two parts, A and B. Tokens represent the embeddings obtained from the segmented words. Subsequently, the input was passed through BERT's encoder, and the output was fed into the classifier to generate the IOB tags. Finally, the data were randomly divided into three datasets for model training: the train set (80%), the validate set (10%), and the test set (10%).

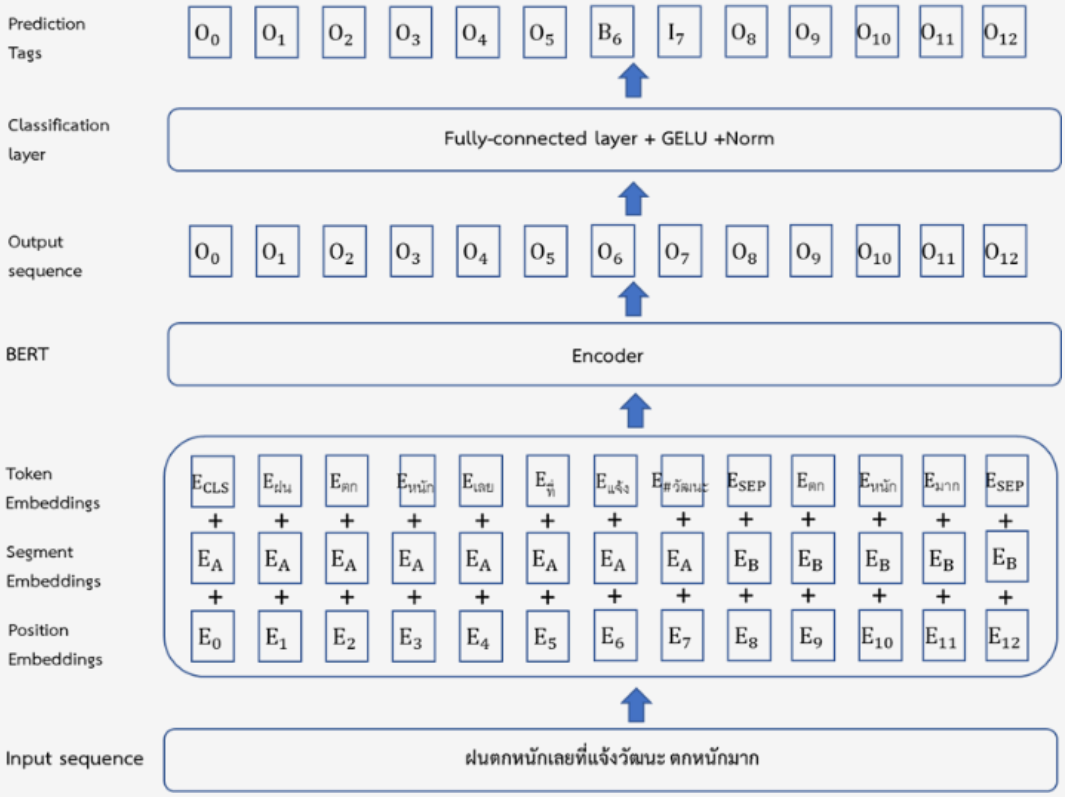

Figure 2: Summarized functions of the employed BERT architecture

4.2 Clustering Algorithm for Geolocation Estimation

As described in Section 2.3, the toponym recognition model was used to extract toponyms from the texts. In cases where the extracted toponyms could not be matched to any coordinates in the database, other toponyms within the same text were used to assist in estimating the geolocation. According to Figure 3, the data from Twitter went through the toponym recognition tool developed in Section 4.1. The next step involved geocoding, where the tool generated geographic coordinate data as output if it could successfully match the toponyms with service providers such as Google's Geocoding API. However, if online service providers could not perform geocoding, clustering algorithms were executed to estimate approximate coordinates. The available clustering models for this task were Topology Words, DBSCAN, K-means, K-medoids, and Agglomerative Clustering.

Once the data was inputted into the algorithm, analyzing whether the target toponyms (words that the Geocoding API could not geocode) appeared in other sentences became imperative. In line with this objective, the study prepared a test dataset containing 100 sets of toponyms. Each set included around 3-20 messages, resulting in a total of 430 messages. Following this process, an analysis was carried out to identify words within geospatial proximity and determine the geospatial boundaries of the target toponyms. Examples are illustrated in Figure 4. To ensure accurate geolocation estimation, potential noise was filtered out from the extracted toponyms using the following clustering algorithms.

4.2.1 K-means

The K-means algorithm is an unsupervised machine-learning technique commonly used for clustering problems. This algorithm partitions objects into K groups and replaces each group with the group means, which serves as the group's centroid and measures the distance of data within the same group [17]. The algorithm operates through the following steps:

1) Determine or randomly select K initial values (groups) and define K initial centers, which are referred to as cluster centers or centroids.
2) Assign all objects to groups by calculating the distance between each data point and the center. The data point closest to the center value is assigned to the corresponding group.
3) Calculate the mean of each group to obtain the new center value.
4) Repeat Steps 2 and 3 until convergence is reached when each group's mean or center point no longer changes. Then, finally, terminate the process.

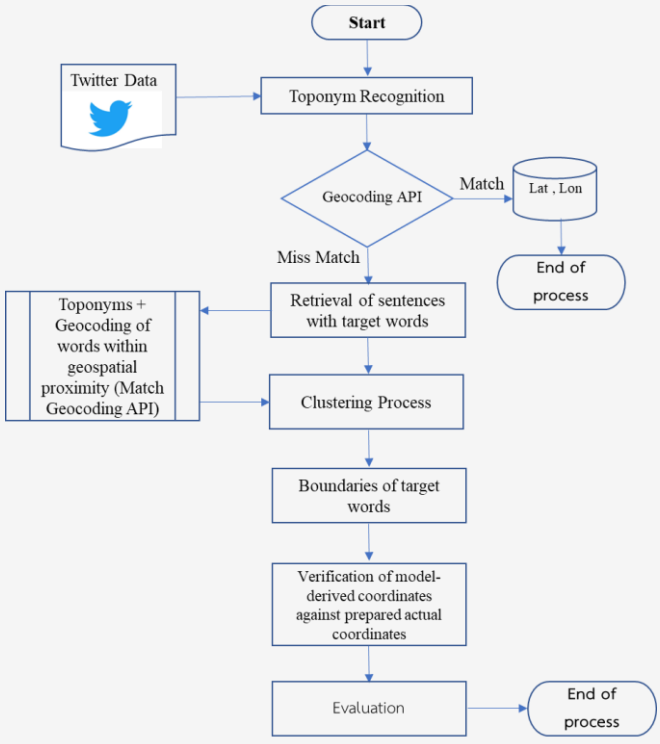

Figure 3: geolocation estimation process

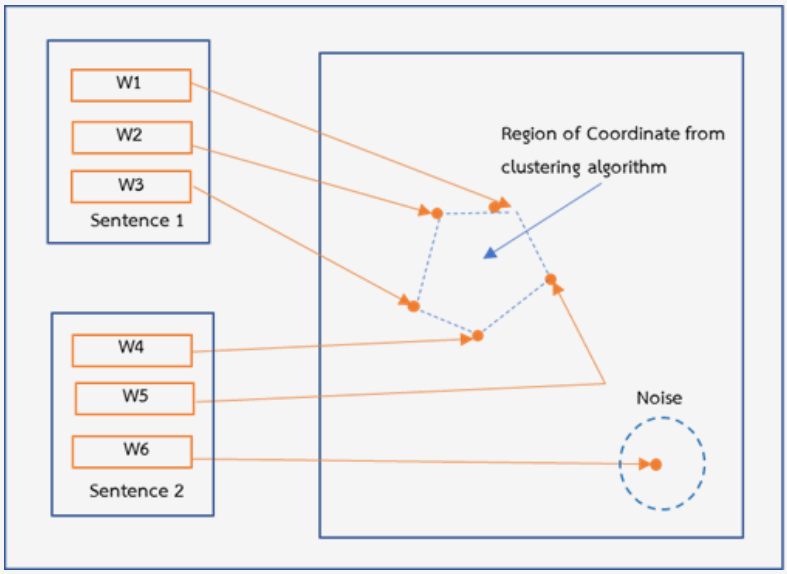

Figure 4: Clustering algorithm for geolocation estimation using toponyms

4.2.2 K-medoids

K-medoids is a clustering technique similar to K-means, but instead of selecting centroids at random from the dataset, it selects medoids, which are actual data points within the dataset, to serve as the center of the clusters [18].

4.2.3 Agglomerative clustering

Agglomerative clustering is a technique that does not require pre-determination of the number of clusters to be formed. Instead, it involves a step-by-step analysis consisting of the following primary steps. Initially, each data point is considered a separate group. For instance, in the case of five toponyms that can refer to geolocations, they can be initially divided into five separate groups.

1) An N x N metric is constructed from the data set, and the distance between each data point is computed.

2) Identify the smallest distance between the data points and group them. Then, improve the metric by selecting the larger of the two groups.

3) The updated data table is retained by keeping the maximum value of the two groups. For instance, if point 1 has a value of 10 and point 2 has a value of 7, the value of 10 is selected.

4) Steps 3-4 are repeated until a single group remains. The final output illustrates the relationships between each data point. Hence, agglomerative clustering is another suitable technique for supporting geolocation estimation [19].

4.2.4 DBSCAN

DBSCAN (density-based spatial clustering of applications with noise) DBSCAN is an unsupervised machine learning segmentation algorithm. This method is considered more robust than other comparable algorithms, such as K-means and hierarchical clustering, as it can handle data clusters with various shapes that are not clearly defined and remove noise data that does not belong to any group. Generally, DBSCAN requires two critical parameters for its operation:

1) Eps represents the maximum radius around a core point, including neighboring data points. These neighboring points are determined based on the parameter MinPts.

2) MinPts is the minimum number of data points required to form a cluster around the center point [20]. Various approaches exist for defining MinPts, but for this study, the approach of Devkota et al., [21] was used, which specifies that MinPts should be equal to the dimension of the data plus one, but not less than three Sarma et al., [22].

In summary, the algorithm functions as follows:

a) Set the value of Minpts to N points.
b) Identify the Core Point X as the point with the closest N neighboring points (including X itself).
c) Move the Core Point X to another point within its Eps radius. If nearby points are within the Eps radius, group them with X in b) to form a cluster.
d) Repeat step c) for all points in the dataset. Points not within the Eps radius of any core point

are classified as noise and are not used to estimate the geolocation.

4.3 Use of Topology Words

Geolocation estimation involves using topology words to determine the location of a place. To illustrate, consider the following sentence: "The head office of the Metropolitan Electricity Authority **is situated in the vicinity** of Khlong Toei, **adjacent to** Rama IV Road." As noted, the sentence contains bolded topology words that indicate the topological relationships between locations. These relationships are assigned weights, such as a weight of 3 for the word "adjacent" and a weight of 1 for the phrase "situated in the vicinity." The weights are assigned using a dictionary-like structure in Python, and the influence of the relationships is shown in detail in Figure 4.

The following pseudocode demonstrates the inner workings of the algorithm in Figure 5. The pseudocode below uses the topology word dictionary, a set of words and their corresponding weights, as denoted by T. The dictionary is defined as follows: {'is at': 3, 'is in ': 3, 'in': 3, 'at': 3, 'on': 3, 'is on ': 3,..., 'in the zone': 1, 'area': 1, 'around': 1, 'surrounding': 1}. G is a list of toponyms and their associated tags, such as [['Bangkok Christian School', 'ACP'], ['Adjacent to', 'GEO'], ['Pramuan Road', 'BSN']] and so on. Finally, the result of the algorithm is a list denoted by R, which contains the processed output.

```
weight_topology_dict = {
   'อยู่':3, 'อยู่ที่':3,'อยู่ใน':3, 'ใน':3,'ที่':3,'บน':3, 'อยู่บน':3, 'ตั้งอยู่':3,
'ที่อยู่':3, 'พิกัด':3, 'ติดกับ':2, 'ใกล้กับ':2, 'ถัดไป':2, 'ต่อกับ':2, 'ถัดจาก':2,'
ใกล้ๆ':2,'ตรงข้าม':2, 'ตรงข้ามกับ':2,'เยื้องๆ':2, 'เยื้องกับ':2,'ใจกลาง':2, 'อยู่
ห่างจาก':1, 'อยู่ห่าง':1,'ห่างจาก':1, 'ห่างไป':1, 'ห่างออกไป':1,'อีกไม่ไกล':1,'
ในเขต':1,'พื้นที่':1, 'บริเวณ':1,'รอบๆ':1 }
weight_topology_dict = {
   'is at': 3, 'is at ': 3, 'is in ': 3, 'in': 3, 'at': 3, 'on': 3, 'is on
': 3, 'is located at': 3, 'coordinates': 3,
   'adjacent to': 2, 'close to': 2, 'next to ': 2, 'connected
with': 2, 'next to ': 2, 'very close to': 2, 'opposite': 2,
'opposite to ': 2, 'slightly next to ': 2, 'at the center of': 2,
   'is far from': 1, 'is far': 1, 'far from': 1, 'further': 1,
'further to': 1, 'not very far': 1, 'in the zone': 1, 'area': 1,
'around': 1, 'surrounding': 1 }
```

Figure 5: Weight values of topology words

```
Requires: T, G, R
Ensure: G <> None
for i ∈ Index of G, g ∈ G :
  bound = length of G
  j = i+1
  k = i+2
  if i <= bound :
    if j < bound & g[j][index of tag] == 'GEO' :
      if g[j][index of tag] <> 'GEO' or 'O' :
        get score from T
      else :
        score = 0
      r = [g[k], score]
      append r To R
end for
Return R
```

Figure 6: Topology words psudocode

4.4 Assessing Model Validity

4.4.1 Assessing the accuracy of the toponym extraction model

This study utilized three matrices, including precision, recall, and overall validity (F1), at the phrase level. This approach enables the assessment of all valid words rather than solely focusing on individual correct words and ensures that the contribution of each valid word is accounted for in the assessment process.

$$Precision = \frac{TP}{TP+FP} \quad \text{Equation 1}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{Equation 2}$$

$$F1 = \frac{2\ x\ (Precision\ x\ Recall)}{Precision\ +\ Recall} \quad \text{Equation 3}$$

where:

TP = The number of toponyms recognized correctly.
FP = The total numbers of toponyms in the system output.
FN = The total numbers of true toponyms in the dataset.

Assessing the performance of models using F1-Token alone may not provide a comprehensive outcome since the task involves extracting complete toponyms from text rather than parts of them. For instance, the sentence "วิวแม่น้ำโขงนครพนมสวยสุดละนี่ชอบมาก" : "The view of Mekong River in Nakhon Phanom is the most beautiful. I like it very much." exemplifies this concern, as depicted in Table 2. From the data presented in Table 2, it is possible to calculate F1-Token as follows: TP = 2, FP = 1, FN = 1, resulting in the precision of 2/(2+1), the recall of 2/(2+1), and the F1-Token value of 2*[(0.67*0.67)/(0.67+0.67)] = 0.34. For F1-Phrase annotations, tags with two tokens are merged into a single tag, meaning the model's answer is considered incorrect if any part of the token is incorrect. For example, if the tokens and tags are (River, B-NAT), (Mekong, I-NAT), and (Nakhon Phanom, B-ADMIN), then the token and tag sequence for the phrase would be Mekong, NAT, and Nakhon Phanom, ADMIN. In Figure 3, for the word Mekong River, the model provided an incorrect answer for the last token, and therefore, the model's answer was deemed incorrect. Using the values from the example above, F1-Phrase can be calculated from TP = 1, FP = 1, and FN = 1. Precision is calculated as 1/(1+1), Recall is calculated as 1/(1+1), and the F1 score is 2*[(0.5*0.5)/(0.5+0.5)] = 0.25. Hence, based on the above example, it is more appropriate to assess performance at the F1-Phrase level since the expected end result of the model is to extract complete toponyms.

4.4.2 Assessing the accuracy of toponym geolocation estimation

RMSE was selected as the metric of this assessment. In addition, intervals were assessed based on the number of points that fell within the radius of the corresponding toponyms. The intervals were divided into four ranges, namely 0.85, 3.39, 8.47, and 10.16 kilometres, to provide a comprehensive assessment. The accuracy of the map was determined at different scales, utilizing the formula outlined by [23] as follows:

> If the scale is larger than 1:20,000, the following calculation should be used to determine ground meters: 0.03333 x scale x 2.54 / 100. However, if the scale is 1:20,000 or smaller, the calculation to use is: 0.02 x scale x 2.54 / 100. For example, at a scale of 1:1,000, the calculation is as follows: 0.03333 x 1000 x 2.54/100 = 0.85 ground meters.

Table 2: Illustration of F1-Phrase identification

Word unit	Annotation	Predicted	parameter
วิว : View	O	B-NAT	FP
แม่น้ำ : River	B-NAT	B-NAT	TP
โขง : Mekhong	I-NAT	O	FN
นครพนม : Nakhon Phanom	B-ADMIN	B-ADMIN	TP
สวย : Beautiful	O	O	
สุดละ : The Most	O	O	
นี่ : I	O	O	
ชอบมาก : Like it very much	O	O	

5. Results

5.1 Toponym Recognition Model Training

The initial phase of this study involved constructing a toponym recognition model based on WangchanBERTa. The training process was conducted under a set of specific environmental conditions, including a seed value of 9, a learning rate of 0.00002, a weight decay of 0.01, and a total of 10 epochs. The training dataset consisted of 17,955 sentences for training, 4,510 sentences for validation, and 5,617 sentences for testing. The results of the training process are shown in Table 3. According to Table 3, the optimal F1 obtained during model training was 0.9188. This result was achieved by utilizing the cosine_with_restarts learning rate scheduler (lr_scheduler_type). In addition, four experiments were conducted with the warm-up ratio suggested by Mishra and Sarawadekar [24]. Furthermore, as demonstrated in Table 4, the WangchanBERTa-based model exhibited the highest overall accuracy (F1) compared to the other models. Table 4 suggested that the BERT model achieved the highest F1 at 0.919 compared to other models. In the context of toponym recognition, additional training from the WangchanBERTa model in this study proved suitable. Table 5 illustrates the F1 values achieved by category. Table 5 presents the results of the fine-tuning process, indicating that the Regional Place (RP) category achieved the highest F1 score of 0.962, while the Monument (MON) category had the lowest F1 score of 0.790, which includes monuments, roundabouts, and clock towers.

Table 3: Training results of the toponym recognition model using BERT architecture

Model	Lr_scheduler_type	Warmup ratio	F1-Phrase
1	-	-	0.9177
2	linear	0.1	0.9169
3	polynomial	0.05	0.9172
4	cosine_with_restarts	0.05	**0.9188**

Table 4: F1-phrase accuracy across models

Index	Model	F1-Phrase
1	CRF PyThaiNLP	0.80
2	CRF Custom feature	0.863
3	LSTM	0.626
4	Bi-LSTM	0.809
5	Bi-LSTM-CRF	0.859
6	WangchanBERTa finetune	**0.919**

Table 5: Accuracy values by topological categories using WangchanBERT

Tag	Description	Precision	Recall	F1-phrase
ACP	Academic place	0.925	0.917	0.921
ADMIN	Admin boundary	0.922	0.934	0.928
BSN	Office building	0.829	0.899	0.863
DEP	Department store	0.930	0.950	0.940
FPLACE	Location outside Thailand	0.883	0.925	0.904
GOV	Government office	0.903	0.873	0.888
HP	Healthcare place	0.892	0.933	0.912
MKT	market	0.941	0.938	0.939
MON	Monument or roundabout	0.842	0.744	0.790
NAT	Natural place	0.817	0.902	0.857
RCT	Recreations, parks, amusement, stadiums	0.909	0.964	0.936
RES	residential	0.842	0.909	0.874
ROAD	Highway, road, alley	0.907	0.927	0.917
RP	Regional place	**0.957**	**0.967**	**0.962**
MKT	market	0.941	0.938	0.939
MON	Monument or roundabout	0.842	0.744	0.790
NAT	Natural place	0.817	0.902	0.857
RT	restaurant	0.889	0.931	0.909
STORE	Store, shops, local shops,	0.819	0.824	0.821
TRAN	Mass transit, train stations, bus stations, piers, ports, etc.	0.871	0.914	0.892
OTHER	Other places	0.933	0.875	0.903

Based on the characteristics of these two topological categories, it was observed that categories such as Admin Boundary (ADMIN), Market (MKT), Academic Place (ACP), and Recreation (RCT) contained features similar to those of the Department Store (DEP) category and could yield similar F1 values. For example, the toponyms of important religious places often contained prefixes such as Wat (วัด) and Masjid (มัสยิด) and suffixes such as Wat, Aram (อาราม), and Wanaram (วนาราม), while markets often used the Talad (ตลาด) prefix. Therefore, the model could accurately extract toponyms from each of these categories.

5.2 Geolocation Estimation

The geospatial estimation results, compared to the reference coordinates obtained from Google Geocoding, were presented in Table 6 using the root mean square error (RMSE) measured in kilometers based on the dataset of 100 sets of toponyms and 430 messages, with 937 toponyms extracted and 48 toponyms filtered out before going through the geospatial estimation process and the subsequent comparison in four ranges. Based on Table 6, the topology words algorithm exhibited the smallest RMSE value and highest accuracy within a buffer of 0.85 km with 78 points, which was similar to the agglomerative clustering algorithm with 60 points. Meanwhile, the K-medoids model had the least accuracy and the highest error, with an RMSE of 3.943 and an accuracy of 48 points at the 0.85 km level. However, determining the location of each point in geolocation estimation presents a challenge since any point could be a starting point, leading to potential spread far apart. As a result, grouping and filtering out outliers can be quite difficult.

6. Discussion

6.1 Case 1

The following illustration serves as proof of concept that the proposed algorithm could produce a near-zero RMSE as demonstrated through toponym words obtained from **Brunch Paradiso**'s dataset. As demonstrated in Figure 7, the processing of messages involves the extraction of toponyms and topology words from the texts, as illustrated in Figure 8. For instance, the name "Brunch Paradiso" was taken from the text and used as a reference to the place without processing any other part of the name. In addition, the FPLACE tag, which indicates that the location is not in Thailand, was also not processed by the algorithm. In geolocation estimation, the coordinates of these four points were used as information for processing.

Table 6: Comparison of accuracy and RMSE values across models

Model	0.85-km buffer	3.39-km buffer	8.47-km buffer	10.16-km buffer	RMSE (km.)
Topology words	**78**	**22**	**0**	**0**	**0.947**
DBSCAN	57	39	4	0	2.211
K-means	58	26	12	4	3.43
K-medoids	48	32	16	4	3.943
Agglomerative clustering	60	29	11	0	2.183

"ไม่บ่อยนักที่กรุงเทพฯ จะมีร้านอาหารที่เน้นเสิร์ฟเมนูบรันช์เป็นหลักให้ได้ไปนั่งเอ็นจอยกัน ครั้งนี้ BKK. ขอแนะนำร้าน Brunch Paradiso ...", "ร้าน Brunch Paradiso ใช้พื้นที่ด้านหน้าของโรงแรม Shama บนถนนเย็นอากาศ ต้อนรับบรันช์เลิฟเวอร์ ที่นี่ยังเป็น Dog Friendly ที่มาสุนัขเข้ามาได้แต่ต้องมีสายจูงหรือรถเข็นเพื่อไม่ให้น้องไปกวนลูกค้าคนอื่น ๆ เราชอบงานอาร์ตที่ใช้ตกแต่ง ส่วนใหญ่เป็นสตอรีของมื้ออาหาร โดยเฉพาะ Art Piece ชิ้นเด่นที่เราสารภาพว่าไม่รู้ว่าเป็นของศิลปินท่านไหน เราขอเรียกว่า "ขนมปังเดินได้" ที่เป็นทั้งภาพวาดและไฟนีออน ตัวเป็นขนมปังขาเป็นขาไก่ขาเป็ด มันสื่อความเป็นบรันช์ได้ดีมาก","สายบรันช์ต้องแวะมาที่ Brunch Paradiso คาเฟ่ในย่านเย็นอากาศ ... "

"Not often in Bangkok will there be a restaurant that serves brunch as the main menu to enjoy. This time, BKK would like to introduce Brunch Paradiso…," "Brunch Paradiso is located in front of the Shama Hotel on Yenakat Road. Welcoming brunch lovers, this place is still dog friendly. You can bring your dog in, but you must have a leash or a stroller so that you don't bother other customers. I love the art decoration, most of which tells the story of the meal. In particular, what is outstanding is the art piece that I must admit that I don't know which artist it belongs to. Let me call it 'Walking Bread.' It is both a painting and a neon light. The body is bread, legs are chicken and duck legs. It conveys brunch very well.," "Brunch lovers must visit Brunch Paradiso, a cafe in Yenakat…."

Figure 7: Sample sentences in case 1

```
[[['ที่', 'GEO'], At
 ['กรุงเทพมหานคร', 'ADMIN'], Bangkok
 ['ที่', 'GEO'], At
 ['brunchparadiso', 'RT'],
 ['ใน', 'GEO'],In
 ['ที่', 'GEO'], At
 ['โรงแรมshama yen-akatbangkok', 'RES'], Hotel shama yen-akat bangkok
 ['เกาะฮ่องกง', 'FPLACE'], Hongkong Island
 ['ใน', 'GEO'],In
 ['ที่', 'GEO']],At
[['ร้าน', 'RT'], Shop
 ['brunchparadiso', 'RT'],
 ['ที่', 'GEO'], At
 ['พื้นที่', 'GEO'], Area
 ['โรงแรมshama', 'RES'], Shama hotel
 ['ที่', 'GEO'], At
 ['บน', 'GEO']], On                    In          At
[['ที่', 'GEO'], ['brunchparadiso', 'RT'], ['ใน', 'GEO'], ['ที่', 'GEO']],
[['ถ.เย็นอากาศ', 'ROAD']]] Yenskat Road
```

Figure 8: Illustration of the extraction of toponyms and topology words from text

Figure 9: Algorithm results on the map

As demonstrated in the example and as illustrated in Figure 8, the message processing involved the extraction of toponyms and topology words from the text. Specifically, the algorithm extracted the name "Brunch Paradiso" from the text to reference the place without processing any other part of the name. Additionally, the FPLACE tag, which indicated that the location was not in Thailand, was not processed by the algorithm. In geolocation estimation, the coordinates of these four points were used as information for processing.

The RMSE results obtained from each algorithm are presented as follows. The Topology Words algorithm had only one point of reference: the Shama-yenakat Hotel in Bangkok. The DBSCAN, K-means, and K-medoids algorithms used all four data points but failed to remove the point that represented the position of Bangkok, which was far from the group of points surrounding the Shama-yenakat Hotel. This led to a high RMSE value of 2.32 km. The Agglomerative Clustering algorithm eliminated the coordinates of Bangkok by grouping three points, and produced the lowest RMSE result of 0.029 km. This was because the algorithm could identify that Brunch Paradiso had a location in the Shama-yenakat Hotel, whereas other algorithms could not filter out such specific information. Finally, the agglomerative clustering algorithm yielded an RMSE of 1.41 km, as Bangkok locations were filtered out, which narrowed the estimated boundaries of the previous algorithm. The map shown in Figure 9 displays the estimated locations obtained from each algorithm, with blue markers representing the locations obtained directly from the toponym recognition tool, green markers representing the estimated locations, and red markers representing the actual locations of the toponym.

"เมื่อวันที่ 12 ธ.ค. เพจ “บูม บาม” ได้โพสต์เรื่องราวของร้านขายของชำที่ชื่อว่าร้าน “จีฉ่อย” สำหรับร้านจีฉ่อย เป็นร้านขายของชำขนาดหนึ่งคูหา ตั้งอยู่หน้าตลาดสามย่าน ถนนพญาไท ตรงข้ามจุฬาลงกรณ์มหาวิทยาลัย ขึ้นชื่อในบรรดานิสิตจุฬาลงกรณ์มหาวิทยาลัย ว่า มีของขายทุกอย่าง และถ้าของไหนไม่มีขายในร้าน จะสามารถมาเอาได้ภายใน 2 วัน ปัจจุบันย้ายร้านไปที่อาคารยูเซ็นเตอร์ 1 ซอยจุฬา 4 ถนนพระราม 4","เมื่อปี 47-48 ผมได้มีโอกาสผ่านอยู่บ่อย ๆ ..... ร้านทำผม ร้านข้าวหน้าเป็ด ร้านโจ๊ก จีฉ่อย ถ้าจำไม่ผิดจะมีร้านทำกุญแจอยู่ด้วย", "ร้านจีฉ่อย ร้านขายของชำในตำนานบนถนนสามย่าน ตึกแถวขนาด 1 คูหาอยู่หน้าตลาดสามย่านเปิด 24 ชม. .....ไม่ยอมขายแต่ก็รู้ทีหลังว่าร้านนี้คือตำนานของเด็กจุฬาก็ที่ไปเลย"

“On December 12, the page ‘Boom Bam’ posted the story of a grocery store called ‘Ji Choi.’ It is a single-block grocery store located in front of Samyan Market on Phayathai Road, opposite Chulalongkorn University. It is famous among Chulalongkorn University students for having everything for sale. If any item is not sold in the store, customers can get it within two days. The shop has since moved to the U Center 1 Building on Soi Chula 4, Rama 4 Road.,” “In 2004-2005, I had the opportunity to pass by often… a hair salon, a duck rice restaurant, a rice porridge restaurant, and Ji Choi. If I’m not mistaken, there is also a locksmith shop.,” “Ji Choi Store is a legendary grocery store on Sam Yan Road. It is a single-block commercial building located in front of Sam Yan Market and is open 24 hours… They would not sell it, but I later found out that it is a legend among Chula students. I was amazed.”

Figure 10: Sample sentences in case 2

Figure 11: Map demonstrating algorithm results when processing more than 10 data points

6.2 Case 2

The following example demonstrates the ability of the proposed algorithms to extract more than 10 toponyms from the text that were not situated inside a department store or an office building, as shown in the example message set from **Ji Choi**'s store geolocation. The above example demonstrated the process of extracting toponyms and topology words from the text. The topology words algorithm had an RMSE of 0.169 kilometres. with four topology word reference points, while the K-means and agglomerative clustering algorithms used six points and had an RMSE of 0.268. The DBSCAN and K-medoids algorithms used eight data points, resulting in an RMSE of 0.532 kilometres, which may have been affected by the larger number of data points. However, the topology words algorithm had a more optimal RMSE value due to the extended range of averaged points, resulting in an error of approximately 0.2 kilometres. The results are presented in Figure 11, where blue markers represent the locations obtained directly from the toponym recognition tool, green markers show the estimated locations from each algorithm, and red markers indicate the actual locations of the toponyms.

7. Conclusion

This study aimed to develop a toponym extraction tool and estimate the geolocations of unknown coordinates. Its initial phase involved developing the tool using the BERT architecture, which was observed to provide a high degree of accuracy, with

a few identified limitations, such as instances where toponyms were extracted incorrectly, with names being incomplete or including additional words such as “within Soi Sukhumvit 24” in the case of “Soi Sukhumvit 24.” To improve and implement the model further, a rule-based filtering approach may be utilized to address these limitations. Alternatively, the model’s overall accuracy could be enhanced through BERT’s encoder and extendedly through the Bi-LSTM-CRF architecture.

In terms of geolocation estimation, this study found that the use of topology words, weighted according to their significance, enhanced the precision of geolocation estimation by reducing noise and improving the accuracy of the resulting geolocations. Future studies may consider assigning weights to different topology words to improve the effectiveness of topological analysis as they may yield different outcomes. Furthermore, the study suggested that filters used for other purposes could be utilized to improve topological analysis. In addition, incorporating other data layers, such as road lines, could further enhance the geolocation estimation process by ensuring that the estimated point is in proximity to the geolocation of the road, potentially resulting in more accurate results.

Acknowledgment

The researchers would like to express their gratitude to the Ratchadapisek Somphot Fund, Chulalongkorn University, for providing the New Lecturer Development Grant and support in the construction of the corpus.

References

[1] Cadorel, L., Blanchi, A. and Tettamanzi, G. B., (2021). Geospatial Knowledge in Housing Advertisements: Capturing and Extracting Spatial Information from Text. *Proceedings of the 11th on Knowledge Capture Conference (K-CAP '21)*, New York, NY, USA, 2021, Vol. 41-48. https://doi.org/10.1145/3460210.3493547.

[2] Kemp, S., (2022). DATAREPORTAL. https://datareportal.com/reports/digital-2021-thailand (accessed 11/10, 2022).

[3] Hahmann, S. and Burghardt, D., (2013). How Much Information is Geospatially Referenced? Networks and Cognition. *International Journal of Geographical Information Science*, Vol. 27(6). 1171-1189. https://doi.org/10.1080/13658816.2012.743664.

[4] Lingad, J., Karimi, S. and Yin, J., (2013). Location Extraction from Disaster-Related Microblogs. *Proceedings of the 22nd International Conference on World Wide Web*, 1017-1020.

[5] Wang, J., Hu, Y. and Joseph, K., (2020). NeuroTPR: A Neuro-Net Toponym Recognition Model for Extracting Locations from Social Media Messages. *Transactions in GIS*, Vol. 24(3), 719-735. https://doi.org/10.1111/tgis.12627.

[6] Zheng, X., Han, J. and Sun, A., (2018). A Survey of Location Prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 30(9). 1652-1671. https://doi.org/10.1109/TKDE.2018.2807840.

[7] Manoruang, D. and Asavasuthirakul, D., (2019). Quality Analysis of Online Geocoding Services for Thai Text Addresses. *Engineering and Applied Science Research*, Vol. 46(2). 86-97. Available: https://ph01.tci-thaijo.org/index.php/easr/article/view/140887.

[8] Chanlekha, H., Kawtrakul, A., Varasrai, P. and Mulasas, I., (2002). *Statistical and Heuristic Rule Based Model for Thai Named Entity Recognition*. Available: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.295.7088.

[9] Chanlekha, H. and Kawtrakul, A., (2004). Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.64.1449.

[10] Tirasaroj, N. and Aroonmanakun, W., (2009). Thai Named Entity Recognition Based on Conditional Random Fields. *Eighth International Symposium on Natural Language Processing*. 216-220. https://doi.org/10.1109/SNLP.2009.5340913.

[11] Thattinaphanich, S. and Prom-On, S., (2019). *Thai Named Entity Recognition Using Bi-LSTM-CRF with Word and Character Representation*. The 4th International Conference on Information Technology (InCIT2019 https://doi.org/10.1109/INCIT.2019.8912091.

[12] Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N. and Nutanong, S., (2021). WangchanBERTa: Pretraining Transformer-Based Thai Language Models. *arXiv*, 2021. [Online]. https://doi.org/10.48550/arXiv.2101.09635.

[13] Xu, D., Cui, P., Zhu, W. and Yang, S., (2014). Find you from your Friends: Graph-based Residence Location Prediction for Users in Social Media. *2014 IEEE International Conference on Multimedia and Expo (ICME)*. 1-6. https://doi.org/10.1109/ICME.2014 .6890 202.

[14] Williams, E., Gray, J. and Dixon, X., (2017). Improving Geolocation of Social Media Posts. *Pervasive and Mobile Computing*, Vol. 36, 68-79. https://doi.org/10.1016/j.pmcj.2016.09.015.

[15] Ramshaw, L. A. and Marcus, M. P., (1999). Text Chunking Using Transformation-Based Learning. *Natural Language Processing Using Very Large Corpora*, S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky Eds. Dordrecht: Springer Netherlands, 157-176.

[16] Devlin, J., Chang, M. W., Lee, K. and Toutanova, K., (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*. https://doi.org/10.48550 /arXiv.1810.04805.

[17] Nanda, A., Barik, R. C. and Bakshi, S., (2023). SSO-RBNN Driven Brain Tumor Classification with Saliency-K-means Segmentation Technique. *Biomedical Signal Processing and Control*, Vol. 81. https://doi.org/10.1016/j.bspc .2022.104356.

[18] Jin, X. and Han, J., (2010). K-Medoids Clustering. *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb (Eds). Boston, MA: Springer US, 2010, 564-565. https://doi.org /10.1007/978-0-387-30164-8_426.

[19] Subasi, A., (2020). Chapter 7 - Clustering Examples. *Practical Machine Learning for Data Analysis Using Python*, A. Subasi Ed.: Academic Press, 465-511.

[20] Kolatch, E., (2001). Clustering Algorithms for Spatial Databases: A Survey. 1-22.

[21] Devkota, B., Miyazaki, H., Witayangkurn, A. and Kim, S. (2019). Using Volunteered Geographic Information and Nighttime Light Remote Sensing Data to Identify Tourism Areas of Interest. *Sustainability*, Vol. 11. https://doi.org/10.3390/su11174718.

[22] Sarma, A., Goyal, P., Kuman, S., Wani, A., Challa, J.S., Islam, S. and Goyal, N., (2019). μDBSCAN: An Exact Scalable DBSCAN Algorithm for Big Data Exploiting Spatial Locality. *IEEE International Conference on Cluster Computing (CLUSTER)*, 23-26 Sept. 2019. 1-11. https://doi.org/10.1109/CLUSTER .2019.8891020.

[23] MAPASYST. (2022). Calculating Your Map Accuracy Using US National Map Accuracy Standards. (Accessed 12/16, 2022).

[24] Mishra, P. and Sarawadekar, K., (2019). Polynomial Learning Rate Policy with Warm Restart for Deep Neural Network. *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, 17-20 Oct. 2019. 2087-2092. https://doi.org/10.1109/TENCON.2019.89294 65.