applied sciences

MDPI

Article

Deep Learning Based Improvement in Overseas Manufacturer Address Quality Using Administrative District Data

Saravit Soeng, Jin-Hyun Bae, Kyung-Hee Lee and Wan-Sup Cho *

Department of Bigdata, Chungbuk National University, Cheongju 28644, Korea
* Correspondence: wscho@cbnu.ac.kr; Tel.: +82-43-261-3636

Abstract: Validating and improving the quality of global address data are important tasks in a modern society where exchanges between countries are due to active Free Trade Agreements (FTAs) and e-commerce. Addresses may be constructed with different systems for each country; therefore, to verify and improve the quality of the address data, it is necessary to understand the address system of each country in advance. In the event of food risk, it is important to identify the administrative district from the address in order to take safety measures, such as predicting the contaminated area by tracking the distribution of food in the area. In this study, we propose a method that applies a deep learning approach to verify and improve the quality of the global address data required for imported food-safety management. The address entered by the user is classified to the administrative division levels of the relevant country and the quality of the address data is verified and improved by converting them into a standardized address. Finally, the results show that the accuracy of the model is found to be approximately 90% and the proposed method is able to verify and evaluate the overseas address data quality significantly.

Keywords: LSTM; RNN; deep learning; address verification; global address verification

Citation: Soeng, S.; Bae, J.-H.; Lee, K.-H.; Cho, W.-S. Deep Learning Based Improvement in Overseas Manufacturer Address Quality Using Administrative District Data. *Appl. Sci.* **2022**, *12*, 11129. https://doi.org/10.3390/app122111129

Academic Editors: Yujin Lim and Hideyuki Takahashi

Received: 30 September 2022
Accepted: 29 October 2022
Published: 2 November 2022

1. Introduction

With the increase in global trade, cross-border e-commerce, and location-based services, the use of verified and accurate addresses has played an important role in business efficiency. The use of high-quality address data saves time and money required for organizational work, increases customer satisfaction, and improves business processes by supporting the accurate delivery of products or services to customers [1,2]. However, each country may have a different address system and the purpose of using an address is different depending on the business; therefore, address data quality verification and improvement are challenging.

The importance of address data quality can vary depending on the field in which the address is used. For example, the accurate delivery of goods is important in the e-commerce system, but in the field of food safety, it is more important to find the correct administrative district from the address or to quickly select optimal area requiring import bans and containment policies in the case of food risk.

Coetzee et al. [3] explained the advantages of address standardization in three ways. The first economic advantage is that it facilitates the exchange of address data by enabling the interoperability of address data. The second social advantage is that some countries have different address systems for each local government; therefore, a standardized address system greatly reduces social confusion. The third advantage of national governance is that addresses play an important role in the performance of public administration tasks, such as elections and censuses.

ISO/TC211, an international standardization organization, established international standards for 「ISO 19160-1: Conceptual Model」 in 2015, 「ISO 19160-4: Postal Address」 in 2017, and 「ISO 19160-3: Address Data Quality」 in 2020 [4]. International

address standardization has been carried out by ISO TC 211 19160 (Geographic Information Division) Working Group 7 (WG7) since 2009 [4]. Conceptual model (ISO 19160-1), address assignment (ISO 19160-2), quality management (ISO 19160-3), international mail (ISO 19160-4), and map notation (ISO 19160-5) are being standardized [4]. In addition, in the era of digital transformation and IoT, it is expanding its scope through standardization of IoT addresses. 「ISO 19160-3: Address Data Quality」 can be the standard for address data accuracy verification [4]. In this standard, address data quality is classified into completeness, logical consistency, positional accuracy, temporal quality, thematic accuracy, and usability element [4].

This study proposed an address verification technique that is useful for improving the quality of address data for foreign food manufacturers in the field of imported food safety. There are approximately 90,000 overseas food manufacturers exporting food to Korea from more than 190 countries [5] and it is necessary to verify the addresses of the manufacturers in each country for efficient business processing. In addition, in the event of a food hazard, it is necessary to promptly recognize the appropriate level of the administrative district from the address of the relevant establishment and take measures to ban or block imports to the relevant area.

Because each country may have a different administrative division system, it may be difficult to recognize the corresponding administrative division from the address. For example, in some countries, the administrative division system has more than four levels, and in small countries, it is composed of less than two levels; therefore, each country must devise a different method. Therefore, a module for verifying addresses and identifying administrative division levels should be developed differently for each country (or group of similar countries) to increase accuracy.

In this study, we verified the addresses of food manufacturers by integrating Google's geocoding service with a deep-learning-based model. Geocoding service separates address components by administrative district level (country, metropolitan city, city, county, etc.) for a given address string and provides latitude and longitude information for the address [6]. However, the address components received from geocoding may not match the actual administrative district database of the country when the address string contains typos or abbreviations (when translating addresses in foreign languages into English, incorrect or abbreviated place names are often used owing to differences in pronunciation, etc.). To overcome the limitations of geocoding, we constructed each country's administrative district database and developed a deep learning model for address verification and standardization. The generated deep learning model is useful in overcoming the limitations of existing geocoding, such as accurately classifying administrative district levels corresponding to address components, even when address strings contain typos or abbreviations. The proposed method can be commonly applied to all countries by using each country's administrative district database. Existing studies mostly verify addresses for specific countries.

Address verification is generally divided into four levels, from the most detailed level: delivery point level, building/house number level, street level, and locality level. This technique is to utilize address verification in food safety management work and aims at verification above the locality level. This is because, in food safety management, it is more important to check the administrative district of the address and take administrative measures, such as export ban, rather than confirming the delivery point when a food risk occurs. The proposed method receives and verifies the English addresses of several countries and does not target the native language addresses of each country.

The proposed method shows more than 90% accuracy in classifying each address component included in the Chinese address dataset by administrative district levels in China. Currently, the Ministry of Food and Drug Safety in Korea is unable to classify it from the address string to the administrative district levels, which is a very advanced result [5]. In addition, spelling errors or abbreviations included in the address are converted to standard terms and the location of the address can be checked on a map, which can be useful

for food-safety management. As results from the verification on manufacturers' addresses of three different countries (China, Japan, and Cambodia), the proposed method is able to verify and evaluate the overseas address data quality significantly.

2. Related Works

In modern society, addresses are becoming a basic infrastructure, going beyond the concept of residence, and are connected to all industries, such as logistics, postal, e-commerce, and location-based industries. The international community is also establishing the address as an international standard to reduce the cost of distribution systems throughout the industry. Recently, the scope of standard enactment has been expanded to address quality, exchange, and maps, and in the IoT era, it is expanding to the standardization of object addresses.

Figure 1 shows an example of address verification. Address verification takes an address string as input and converts it into a correct address string through parsing, formatting, standardization, geocoding, and verification processes (refer to Figure 1) [7].

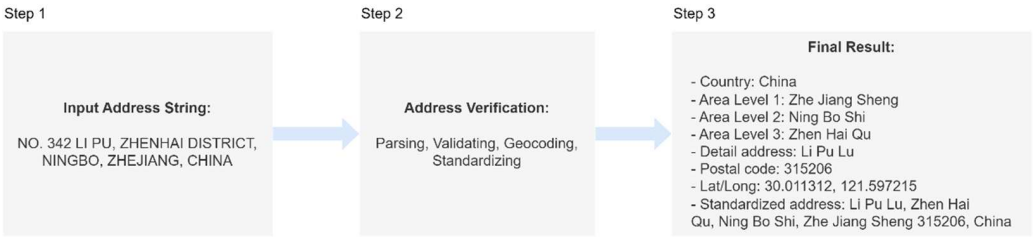

Figure 1. Address verification [7].

Christen et al. [8] detailed address tags, which were constructed based on the Australian National Postal Address Guidelines and the Australian Geocoded National Address File (G-NAF). They also presented an automated approach for address cleaning and standardization using a probabilistic hidden Markov model (HMM). Figure 2 delivers an example of the address tag [8].

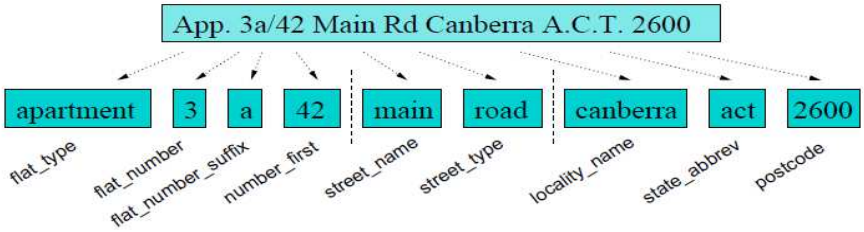

Figure 2. Detailed Australian address tag—an example.

Abid et al. [9] proposed Deepparse, a deep-learning-based multinational address parsing library that can be applied to non-standard address problems, named entity recognition (NER) problems. The library suggested an improved generalization technique for non-standard address data and a solution to the class mixing and entity name recognition problems.

Sharma et al. [10] presented a machine-learning-based technique for address parsing using a neural network. Similarly, Delil et al. [11] used a deep learning approach to apply to address parsing tasks in their research by utilizing the convolutional neural network. Another study, by Li et al. [12], offered a Hidden-Markov-Model-based approach to parse the addresses by building the models based on the synthetic training data.

Min et al. [13] proposed a method for detecting location information of an extended concept that includes administrative districts, names of institutions, libraries, and movie theaters in text data that are not geo-tagged. Using unstructured text data extracted from news, articles, blogs, and social media, a deep learning model based on labeling, word

embedding, and attention was used to create a binary classifier and predict whether place information was included.

Matci and Avdan [14] proposed a technique for standardizing addresses using natural language processing for improving geocoding results. The research addressed problems, such as inaccurate numbering systems, misspellings, the use of abbreviations, and a lack of data by way of a standardization process that decomposes addresses used as input data in geocoding by identifying spelling mistakes and abbreviations and reorganizing the address via natural language processing.

Guermazi et al. [15] provided a RoBERTa-based approach to validate the address. The research utilized a two-step address verification approach consisting of standardization and classification. Both steps depended on RoBERTa, which is a pre-trained language model. After the experiments on the real dataset, the result showed the effectiveness of the proposed technique compared to the alternative approaches.

The research by Xi et al. [16] suggested an original joint learning strategy based on the hash map principle and word frequency theory to standardize Chinese non-standard building addresses. The proposed research was to address the issues of using traditional methods based on string matching that struggles to meet the task requirements because of the substantial number of non-standard building addresses and the semantic ambiguity of addresses stated in the Chinese natural language.

Lu et al. [17] suggested a way for standardizing addresses for Chinese addresses based on the seq2seq model. The research makes use of attention mechanisms to assess the relative importance of the various components of the address and the Gated Recurrent Unit (GRU) to learn the intrinsic link in the Chinese address. Without requiring extra information, such as a standard address database or a geological element table, the method can fill in the blanks in the administrative address information and fix incorrect address information. Other scholars [18–20] also conducted in-depth research to carry out the difficulty of address standardization with various aspects.

Another study by Lee et al. [21] proposed an address geocoding system that makes use of machine learning to improve the street-based address-matching method. Address parsing, address matching, and address locating are the three modules that make up the developed address geocoding algorithm. The input addresses are divided into components of the street-based system using a regex-based parsing approach. This paper provides a method to integrate similarity measures for address matching in order to enhance performance.

A study by Cebecı et al. [22] aimed to create a system that matches free-text address data with conventional addresses using the Support Vector Machines method. A free text address's resemblance to a standard address is expressed as a numerical value by a model that was trained using categorized data. The research confirmed that by applying the proposed system to the free-text addresses generated from 250,000 addresses, the system could achieve a matching accuracy of over 81%.

Xu et al. [23] introduced an address-matching technique based on deep transfer learning to identify semantic similarities between various addresses. The proposed study firstly pre-trained the address corpus to learn address contexts unsupervised and then created a labeled address-matching dataset that enables the matching problem to be transformed into a binary classification prediction problem by utilizing the specific geographic feature. The study finally applied the fine-tuning technique by using the address-matching dataset to build the classification model. The results showed that their model performed at the highest level, with precision, recall, and F1 score above 0.98. Another similar study by Shan et al. [24] proposed a geographical address representation learning for address matching. The proposed research learned the geographical semantic representations for address strings by obtaining rich contexts for addresses from the Web via search engines. The study utilized an encoder–decoder architecture to learn semantic vector representation for each address string along with the attention mechanism and then constructed a large graph from the corpus containing address elements and addresses as nodes. Word co-occurrence data are used to build the edges between nodes in order to develop embedding represen-

tations for every node on the graph. The study claimed that their proposed method outperformed the existing methods in terms of precision and recall.

Meanwhile, Lin et al. [25] also applied a deep-learning-based architecture for semantic address matching in their research. The study trained the word2vec model to convert the address data into their respective vector representations and then applied the enhanced sequential inference model to make inferences to determine the matching of two addresses. The research used real-world address data from the Shenzhen Address Database to evaluate the proposed method. The result showed that the proposed method gained a higher matching accuracy for unstructured address data, with its precision, recall, and F1 score up to 0.97. Another study [26–31] from related scholars also proposed different methodologies to deal with address-matching problems in various contexts by utilizing the most advanced technologies, such as deep learning and machine learning.

All studies significantly proved the results in their own way. However, the previous study mostly focused on the techniques for parsing, matching, and standardizing the address. Some research specifically focused on a specific country. Anyway, it is clear that working with addresses is one of the most interesting and complicated tasks to undertake. In our research, we offer more approaches rather than parsing, matching, and standardizing the address. This study focuses more on tasks of classifying the address components parsed from the full address string into its correct administrative area levels based on each country. Further, the research provides the technique to verify and evaluate the quality of an address based on the results from the classification model of address components that relies on a deep learning approach. The purpose of the research is to work on many available countries rather than a specific country. Further, in this research, as we have our limitations and scope, we first decide to work on three different countries, China, Japan, and Cambodia, and to extend more in the future.

3. Deep-Learning-Based Address Verification Technique

The classification of address components at the administrative district level is necessary when implementing import restrictions or containment measures in food-safety management. This section introduces a deep-learning-based classification model that divides an address string into address components and classifies each component into an appropriate administrative district level.

We developed an address verification program using a deep learning technique (Figure 3). In this research, we use the deep learning technique for the text (address components) classification task as the deep learning technique may outperform other traditional machine learning algorithms in terms of performance, accuracy, and adaptability [32–36]. First, a Long Short-Term Memory (LSTM)-based multi-class classification model was created after labeling and character embedding on country-specific administrative district data. Next, in the actual address verification step, Google geocoding is used to generate a string for each administrative district (country, Level 1, Level 2, Level 3) from the address string and then input it into the deep learning model. Finally, the accuracy of the address quality is calculated and the address is improved by converting it into a standardized address format.

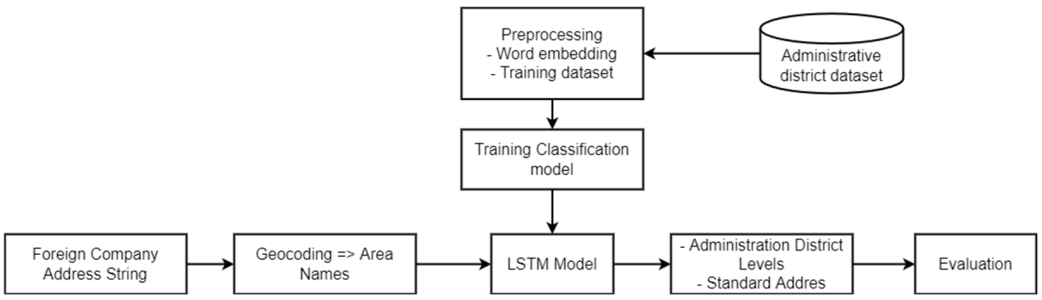

Figure 3. Proposed methodology flow.

3.1. Dataset Preprocessing

The first step is to collect each country's administrative district dataset and create a training dataset for the deep learning model. Each country's administrative district dataset is created manually using each country's address system database that is collected from Postcode Query website [37]. Table 1 shows an example of the administrative district dataset for Korea consisting of country, level 1, level 2, and level 3 (level 3 is the requirement for Korea Food and Drug Safety Administration).

Table 1. Administrative district sample dataset (Korea).

Country	Level_1	lv1_Division	Level_2	lv2_Division	Level_3	lv3_Division
Korea	Chungcheongbuk	Do	Cheongju	Shi	Seowon	Gu
Korea	Gyeonggi	Do	Anseong	Shi	Bogae	Myeon

Table 2 shows the training dataset created using Table 1. It is a table consisting of the regional name of each country and the administrative division level of that region.

Table 2. Sample training dataset.

Region Name	Label
Korea	country
Chungcheongbuk	level_1
Cheongju	level_2
Seowon	level_3

3.2. Deep Learning Model

Figure 4 shows the generation of a deep learning model [38–41]: word embeddings, LSTM layers, and dense layers.

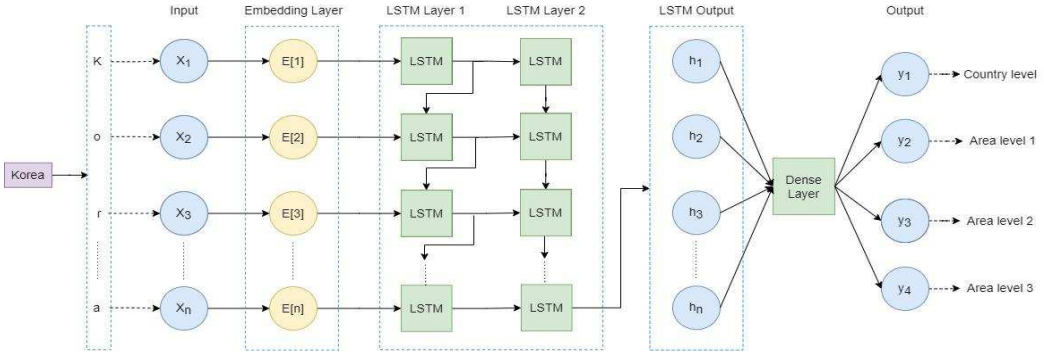

Figure 4. Creation of a deep-learning-based address learning model for address validation.

A text vectorization technique is used to vectorize address elements before using them in a deep learning model [39]. Text vectorization can be divided into two levels: word and character levels. In this study, the character level was selected because learning is conducted about the components of the address rather than the entire string constituting the address. Thus, address elements, including spelling errors, can be classified at the appropriate administrative district level according to the degree of similarity. In addition, even when an abbreviation or code is used, accuracy can be increased by including it in the training dataset.

The Long Short-Term Memory (LSTM) layer is a special type of recurrent neural network that can learn order dependence in sequence prediction or classification problems [41–43]. In this research, the LSTM model consists of two LSTM layers with 128 hidden units in each LSTM cell.

The dense layer applies the softmax activation function to divide the classification results into four different classes: country level, region level 1, region level 2, and region level 3. For example, if "Korea" is input into the LSTM model, "Country level" is output from the softmax function.

3.3. Address Verification Process

The generated LSTM model was used to validate the address strings in the field and classify them at the administrative level. When an address string for a foreign company is input, it is divided into address components using Google geocoding. When each address component is input into the LSTM model, the corresponding administrative district level is output. The accuracy of the address is calculated and, if necessary, is converted into a standard address format. Figure 5 illustrates the end-to-end address verification flow.

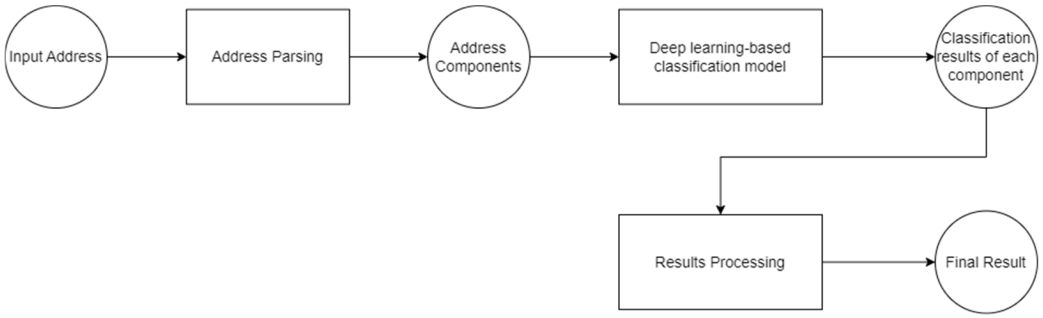

Figure 5. End-to-end address verification process.

3.4. Evaluation Metrics

The evaluation metrics are utilized to measure and summarize the quality of the trained classifier when tested with the unobserved data [44]. Generally, the overall accuracy is the most commonly used metric for classification models, but the only accuracy is not enough to assess the model for some reasons [45]. It requires other metrics to check and verify collectively to assure that the model works properly.

Principally, true positives (TP) and true negatives (TN) are described as outcomes of the positive class and negative class, respectively, correctly classified by the model. Meanwhile, false positive (FP) and false negative (FN) are the outcomes of positive and negative classes classified incorrectly. Accordingly, the overall accuracy is calculated by Equation (1).

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \tag{1}$$

The precision, recall, and F1 score are calculated by Equations (2)–(4)

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

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{3}$$

$$\text{F1} - \text{Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{4}$$

The additional significant metric is the area under the curve (AUC) of the ROC (Receiver Operating Characteristic) curve. The ROC curve provides a visual implement for checking the ability of the prediction model to correctly classify positive cases and negative cases that were incorrectly classified [45,46].

3.5. Model Evaluation

Among foreign companies exporting food to Korea, Chinese companies are the most common; therefore, we verify the accuracy of the proposed system with addresses in several countries, including China.

Here, we tested how well the classification model verified addresses of Chinese companies. The dataset consists of 35,373 address components and it is structured as shown in Table 2. Among these, 28,662 components were used for the training dataset and the remaining 7075 components were used as the test dataset. Note that address components corresponding to the country ~level 3 have many duplicate values (since the country~level 3 has the same value, but many different addresses occur below level 4). Therefore, the test dataset includes almost all address components in China and the training dataset is sufficient for model training.

Table 3 lists the results of evaluating the performance of the model using the confusion matrix method and Table 4 lists the precision, recall, and F1 scores for each class.

Table 3. Confusion matrix.

Actual \ Predicted	Country	Level_1	Level_2	Level_3
Country	1788	0	0	0
Level_1	0	1802	17	0
Level_2	0	18	1740	7
Level_3	0	4	117	1582

Table 4. Precision, Recall, and F1-Score.

Class	Precision	Recall	F1-Score
Country	100.00%	100.00%	100.00%
Level_1	98.79%	99.07%	98.93%
Level_2	92.85%	98.58%	95.63%
Level_3	99.56%	92.89%	96.11%

Based on the model evaluation results listed in Tables 3 and 4, the predictive model showed a score of 90% or higher in precision, recall, and F1 scores for each administrative district level, with an overall accuracy of 97.70%. Based on these results, the model is judged to be appropriate and accurate for classifying the address components into the corresponding administrative division level. In this research, the ROC curve of model evaluation is also reported in Figure 6.

The previous step is the process of checking the accuracy of each administrative district level generated by the classification model for a given address. The classification results are not the final objective of our research. The average accuracy of the results of classifying all addresses by administrative district level was obtained and scored. For example, assuming that the address components are classified into four components (Country, Level 1, Level 2, and Level 3) and each has an accuracy of 99%, 90%, 90%, and 95%, respectively, the average accuracy is (99 + 90 + 90 + 95)/4 = 93.5%, which is determined to be the address quality accuracy. Depending on the domain or country used, the lower limit of acceptable address quality accuracy may be set.

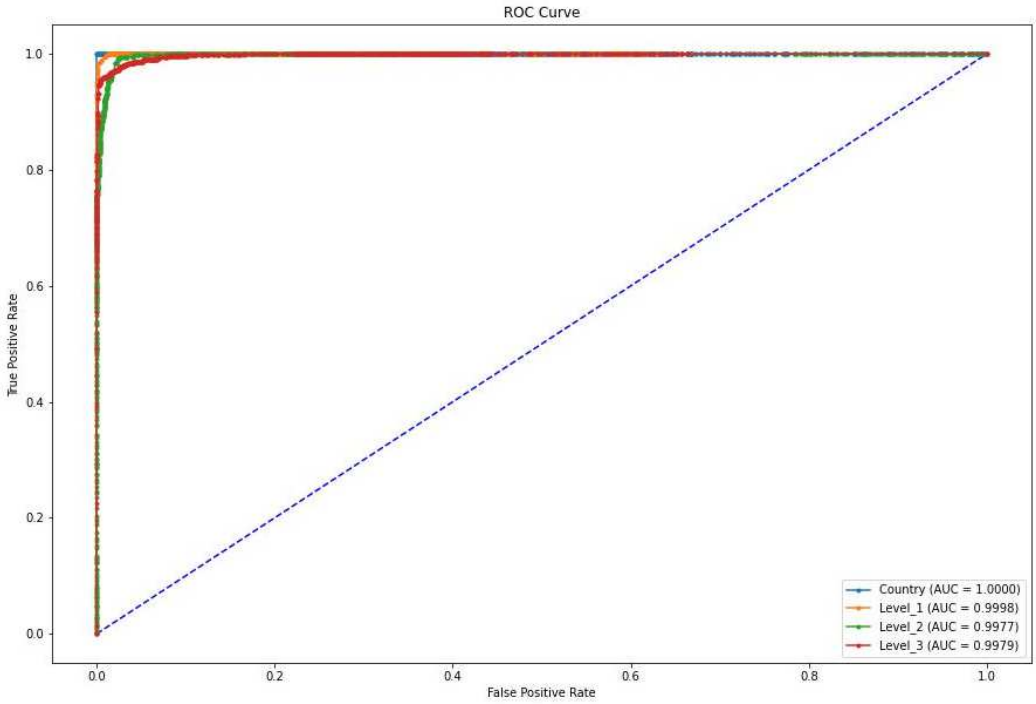

Figure 6. The Receiver Operating Characteristics (ROC) Curve.

4. Use of Address Verification System in Actual Business

In this section, we apply the proposed system to validate overseas food manufacturing companies in real business.

4.1. Experimental Results

Here, addresses from three countries (Cambodia, China, and Japan) were collected from the imported food information portal of the Ministry of Food and Drug Safety of Korea [5]. Table 5 lists the number of manufacturers' addresses in each country. Because of differences in the address systems of the countries, a separate classification model was constructed for each country. The address quality accuracy results for each country are listed in Table 6.

Table 5. Number of manufacturing addresses by country.

Country	Number of Addresses
Cambodia	89
China	29,532
Japan	6519

Table 6. Average accuracy of address quality of manufacturers by country.

Country	Average Accuracy
Cambodia	77.30%
China	87.31%
Japan	89.77%

The results from the experiments are verified and evaluated by using the proposed method in this research. Based on the results from Table 6, it can be inferred that Cambodian food manufacturers' address quality is not up to the mark, as the average address accuracy is just 77.30%. Meanwhile, it appears that the food manufacturers in China and Japan provide good address quality, since the average accuracy of addresses is close to 90%, which is considered to be a good result. Hence, the address quality accuracy describes how

good the address is. If an address receives a low accuracy score, it is because it is poorly formatted, misspelled, informal, non-standard, shortened, or uses slang; otherwise, it receives a high accuracy score.

4.2. Web-Based Address Verification System

A web-based address verification system is implemented so that users who require address verification can conveniently verify the addresses on the web. The system consists of a simple interface that allows the user to select a specific country and enter an address. The system can classify the address by administrative district and verify the accuracy. Figure 7 shows the web-based user interface.

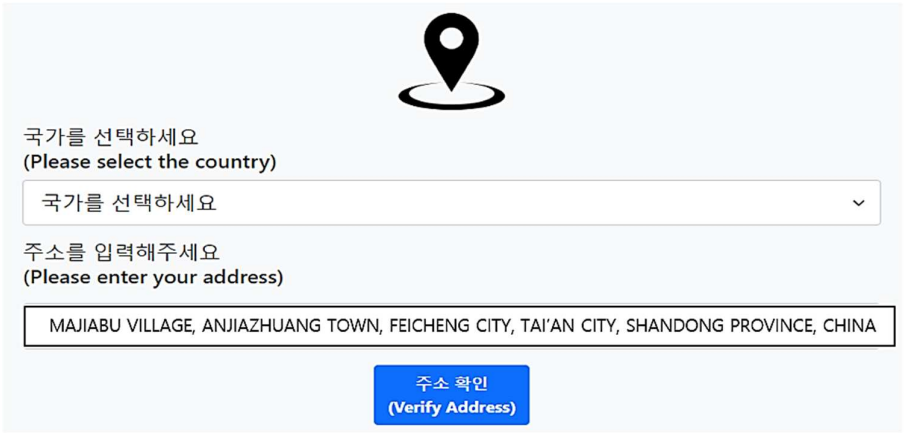

Figure 7. Web-based verification system user interface.

Figure 8 shows the verification result for the address of a Chinese manufacturer using a web-based address verification system. A given address was predicted to be 98.38% correct and the components of the address were segmented at the district level. Thus, the accuracy of a given address can be improved and, simultaneously, it can be converted into a standard address format. In addition, a map corresponding to the entered address is provided on the web using the Google Map service (Figure 9).

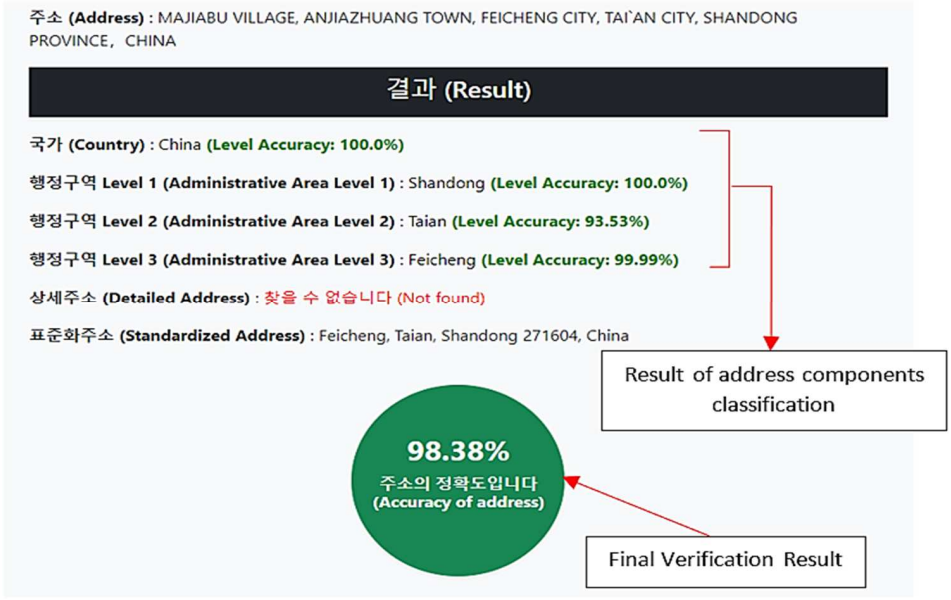

Figure 8. Verification result for a Chinese business address.

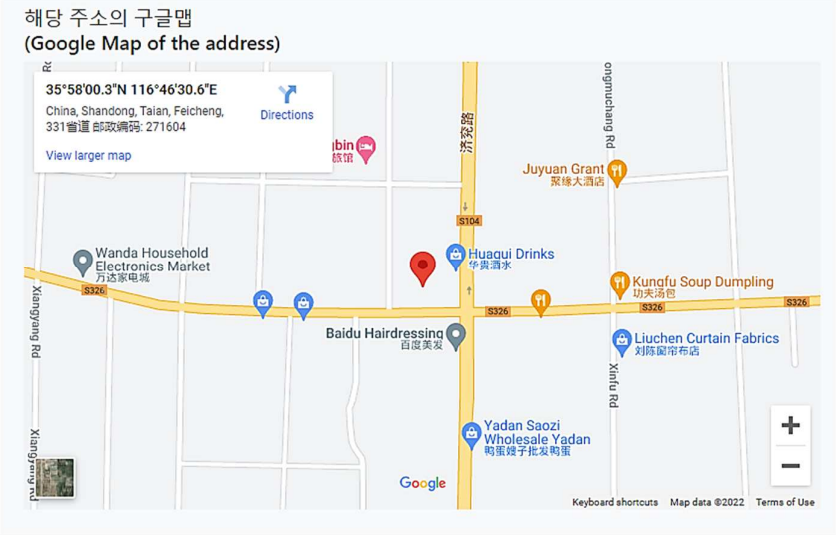

Figure 9. Location for the manufacturer address provided by Google Map service.

5. Discussion

Most of the existing studies deal with address verification processes, such as parsing, matching, and standardizing for a specific context or country, but the proposed method in our research is designed to apply to various countries based on deep learning classification models using each country's administrative district data. The deep learning model is able to determine the level of administrative districts from a given address string and through this, the accuracy of the given address is calculated and the quality is improved by converting it into a standardized address. In addition, it is able to overcome the limitations of existing geocoding services that cannot accurately identify the administrative district level corresponding to the entered address of the manufacturers.

Regarding the experimental results, the proposed method is able to verify and evaluate the manufacturers' addresses significantly and such tasks have not been completed yet in prior research. Prior studies did not primarily concentrate on categorizing the address components to their administrative area level and assessing the quality of the address; however, our study is able to complete such tasks. In this proposed paper, the examination and verification of the manufacturers' addresses of nations that export food to Korea are the exclusive focus of our research. Based on the results from the evaluation of three different countries, such as China, Japan, and Cambodia, the results were expressed differently based on the quality of addresses for each country. Accordingly, it is supposed that the quality of manufacturers' addresses from Cambodia is still insufficient as it could only achieve an average address accuracy of 77.30%, while China and Japan achieved a good score in average accuracy of nearly 90%. That means the addresses of manufacturers from Cambodia can be incorrectly formatted, misspelled, informal, unconventional, abbreviated, or slang, while manufacturers from China and Japan deliver the addresses that meet the standard. Anyway, it should be noted that, unlike China and Japan, in Cambodia, the number of food manufacturers registered with the Ministry of Food and Drug Safety in Korea is only 89 companies and these are used for the evaluation.

Due to the limitation and scope, we can only focus on three countries in this study and hope to extend more in future research.

6. Conclusions

This research introduces a technique designed to parse and verify address quality correctly and efficiently. The proposed model architecture is based on a deep learning type of recurrent neural network. The model is built for classifying the address components that have been parsed from the full address for verification. Our technique does not use the full address. Instead, we work on each address component parsed from the full address string.

In this study, a deep learning model that verifies the addresses of foreign food manufacturers in the field of food safety was constructed and its quality accuracy was evaluated. In the field of food safety, the accuracy of the address quality itself is important, but recognizing the administrative district to which the company belongs for the address string is important for a prompt response in case of harm. Administrative district datasets for each country were constructed manually based on the open-address data from Postcode Query website for each country to create the deep learning classification model. A classification model was distinctively created for each country in consideration of the different address systems. When the address string is input, the address is parsed into address components using Google geocoding service and the deep learning model classifies each address component into an appropriate level of administrative district level. In addition, the input address in this process is also converted into a standardized address. After building and evaluating a model based on the Chinese address data, the accuracy of the classification model is greater than 90% and its precision, recall, and F1 score are all above 90%.

For the experiments, the manufacturers' addresses of three countries, including China, Japan, and Cambodia, were examined and evaluated. Based on the examination of three distinct nations, the experimental findings demonstrate that the suggested method is significantly capable of verifying and evaluating the quality of overseas address data. The results show that the address data quality accuracy was 87.31%, 89.77%, and 77.30% for China, Japan, and Cambodia, respectively.

By configuring and preprocessing the training dataset well, it is possible to overcome problems, such as errors, abbreviations, and synonyms, due to differences in English pronunciation in address, which are known to be difficult in address verification. The deep-learning-based technique in this research is also able to assist any organization or business to build their own address quality verification system without depending on a commercial service.

However, in this study, there is a limitation in terms of scope of research. We only focused on three different overseas countries, such as China, Japan, and Cambodia. The research is expected to extend to more countries in the future. The research can also be continued in order to improve the performance and accuracy of the model. The technique in the proposed method, such as address parsing, will be considered to be updated. Currently, we utilize the service of Google that offers the limitation of the usage. Therefore, we plan to make our own way in parsing the full address string into components.

Author Contributions: Conceptualization, methodology and software, K.-H.L.; software, validation, and investigation, S.S.; writing—original draft preparation, W.-S.C.; formal analysis, J.-H.B.; supervision and project administration, K.-H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant (21163MFDS517-1) from the Ministry of Food and Drug Safety of South Korea in 2022.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The administrative district data is publicly available at https://www.postcodequery.com (accessed on 22 May 2022) and overseas manufacturer address data is available at https://impfood.mfds.go.kr/CFCCC01F02/getList2 (accessed on 19 October 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Address Verification. Available online: https://www.loqate.com/resources/blog/address-verification/ (accessed on 19 June 2022).
2. Why is Address Verification Important? Available online: https://www.smartystreets.com/articles/why-is-address-verification-important (accessed on 19 June 2022).

3. Coetzee, S.; Cooper, A.K. Value of addresses to the economy, society and governance—A South African perspective. In Proceedings of the 45th Annual Conference of the Urban and Regional Information Systems Association (URISA), Washington, DC, USA, 20–23 August 2007.
4. Addressing—Part 3: Address Data Quality. Available online: https://www.iso.org/standard/71247.html (accessed on 27 September 2022).
5. Imported Food Information Maru. Available online: https://impfood.mfds.go.kr/CFCCC01F02/ (accessed on 19 October 2022).
6. Geocoding API. Available online: https://developers.google.com/maps/documentation/geocoding/overview (accessed on 20 June 2022).
7. Address Verification. Available online: https://www.melissa.com/address-verification (accessed on 18 October 2022).
8. Christen, P.; Belacic, D. Automated probabilistic address standardisation and verification. In Proceedings of the Australasian Data Mining Conference, Sydney, Australia, 5–6 December 2005.
9. Abid, N.; Hasan, A.U.; Shafait, F. DeepParse: A Trainable Postal Address Parser. In Proceedings of the 2018 Digital Image Computing: Techniques and Applications (DICTA), Canberra, Australia, 10–13 December 2018; pp. 1–8.
10. Sharma, S.; Ratti, R.; Arora, I.; Solanki, A.; Bhatt, G. Automated Parsing of Geographical Addresses: A Multilayer Feedforward Neural Network Based Approach. In Proceedings of the 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 31 January–2 February 2018; pp. 123–130.
11. Delil, S.; Kuyumcu, B.; Aksakallı, C.; Akçıra, İ.S. Parsing Address Texts with Deep Learning Method. In Proceedings of the 2020 28th Signal Processing and Communications Applications Conference (SIU), Gaziantep, Turkey, 5–7 October 2020; pp. 1–4.
12. Li, X.; Kardes, H.; Wang, X.; Sun, A. Hmm-based address parsing with massive synthetic training data generation. In Proceedings of the 4th International Workshop on Location and the Web, Shanghai, China, 3 November 2014; pp. 33–36.
13. Min, K.; Song, J.; Yu, K.; Kim, J. A Method for Detecting Location Information using Attention-based Deep Learning Model and Word Embedding. *J. Korean Soc. Geospat. Inf. Sci.* **2019**, *27*, 33–39. [CrossRef]
14. Küçük Matci, D.; Avdan, U. Address standardization using the natural language process for improving geocoding results. *Comput. Environ. Urban Syst.* **2018**, *70*, 1–8. [CrossRef]
15. Guermazi, Y.; Sellami, S.; Boucelma, O. A RoBERTa Based Approach for Address Validation. In *New Trends in Database and Information Systems, Communications in Computer and Information Science*; Springer: Cham, Switzerland, 2022; pp. 157–166.
16. Xi, X.F.; Wang, L.; Zou, E.; Zeng, C.; Fu, B. Joint Learning for Non-standard Chinese Building Address Standardization. In Proceedings of the 2018 IEEE International Smart Cities Conference (ISC2), Kansas City, MO, USA, 16–19 September 2018; pp. 1–8.
17. Lu, Y.; Liu, H.; Zhou, Y. Chinese Address Standardization Based on seq2seq Model. In Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems, Bangkok, Thailand, 23–25 November 2019; pp. 1–5.
18. Munjas, I.; Batanović, V. US address classification based on text processing and machine learning. In Proceedings of the 2021 29th Telecommunications Forum (TELFOR), Belgrade, Serbia, 23–24 November 2021; pp. 1–4.
19. Cao, H.N.; Tran, V.T. Deep neural network based learning to rank for address standardization. In Proceedings of the 2021 RIVF International Conference on Computing and Communication Technologies (RIVF), Hanoi, Vietnam, 19–21 August 2021; pp. 1–6.
20. Luo, A.; Liu, J.; Li, P.; Wang, Y.; Xu, S. Chinese address standardisation of POIs based on GRU and spatial correlation and applied in multi-source emergency events fusion. *Int. J. Image Data Fusion* **2021**, *12*, 319–334. [CrossRef]
21. Lee, K.; Claridades, A.R.C.; Lee, J. Improving a Street-Based Geocoding Algorithm Using Machine Learning Techniques. *Appl. Sci.* **2020**, *10*, 5628. [CrossRef]
22. Cebecı, S.; Özyılmaz, M.; İnce, G. Automatic Standardization System for Free Text Addresses. In Proceedings of the 2019 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey, 24–26 April 2019; pp. 1–4.
23. Xu, L.; Mao, R.; Zhang, C.; Wang, Y.; Zheng, X.; Xue, X.; Xia, F. Deep Transfer Learning Model for Semantic Address Matching. *Appl. Sci.* **2022**, *12*, 10110. [CrossRef]
24. Shan, S.; Li, Z.; Yang, Q.; Liu, A.; Zhao, L.; Liu, G.; Chen, Z. Geographical address representation learning for address matching. *World Wide Web* **2020**, *23*, 2005–2022. [CrossRef]
25. Lin, Y.; Kang, M.; Wu, Y.; Du, Q.; Liu, T. A deep learning architecture for semantic address matching. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 559–576. [CrossRef]
26. Comber, S.; Arribas-Bel, D. Machine learning innovations in address matching: A practical comparison of word2vec and CRFs. *Trans. GIS* **2019**, *23*, 334–348. [CrossRef]
27. Cheng, R.; Liao, J.; Chen, J. Quickly locating POIs in large datasets from descriptions based on improved address matching and compact qualitative representations. *Trans. GIS* **2022**, *26*, 129–154. [CrossRef]
28. Zhang, H.; Ren, F.; Li, H.; Yang, R.; Zhang, S.; Du, Q. Recognition Method of New Address Elements in Chinese Address Matching Based on Deep Learning. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 745. [CrossRef]
29. Tian, Q.; Ren, F.; Hu, T.; Liu, J.; Li, R.; Du, Q. Using an Optimized Chinese Address Matching Method to Develop a Geocoding Service: A Case Study of Shenzhen, China. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 65. [CrossRef]
30. Koumarelas, I.; Kroschk, A.; Mosley, C.; Naumann, F. Experience: Enhancing address matching with geocoding and similarity measure selection. *J. Data Inf. Qual.* **2018**, *10*, 1–16. [CrossRef]
31. Cruz, P.; Vanneschi, L.; Painho, M.; Rita, P. Automatic Identification of Addresses: A Systematic Literature Review. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 11. [CrossRef]

32. Kamath, C.N.; Bukhari, S.S.; Dengel, A. Comparative Study between Traditional Machine Learning and Deep Learning Approaches for Text Classification. In Proceedings of the ACM Symposium on Document Engineering 2018, Halifax, NS, Canada, 28–31 August 2018; p. 14.
33. Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P.S.; He, L. A survey on text classification: From shallow to deep learning. *arXiv* **2021**, arXiv:2008.00364 2020.
34. Akpatsa, S.K.; Li, X.; Lei, H. A survey and future perspectives of hybrid deep learning models for text classification. In Proceedings of the International Conference on Artificial Intelligence and Security, Dublin, Ireland, 19–23 July 2021; pp. 358–369.
35. Chen, C.-W.; Tseng, S.-P.; Kuan, T.-W.; Wang, J.-F. Outpatient text classification using attention-based bidirectional LSTM for robot-assisted servicing in hospital. *Information* **2020**, *11*, 106. [CrossRef]
36. Semberecki, P.; Maciejewski, H. Deep learning methods for subject text classification of articles. In Proceedings of the 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, Czech Republic, 3–6 September 2017; pp. 357–360.
37. Postcode Query. Available online: http://www.postcodequery.com/ (accessed on 22 May 2022).
38. Text Classification with an RNN. Available online: https://www.tensorflow.org/text/tutorials/text_classification_rnn (accessed on 21 June 2022).
39. How to Prepare Text Data for Deep Learning with Keras. Available online: https://machinelearningmastery.com/prepare-text-data-deep-learning-keras/ (accessed on 20 June 2022).
40. How to Use Word Embedding Layers for Deep Learning with Keras. Available online: https://machinelearningmastery.com/use-word-embedding-layers-deep-learning-keras/ (accessed on 21 June 2022).
41. A Gentle Introduction to Long Short-Term Memory. Available online: https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/ (accessed on 21 June 2022).
42. Graves, A. Long short-term memory. In Supervised Sequence Labelling with Recurrent Neural Networks. Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.
43. Van Houdt, G.; Mosquera, C.; Nápoles, G. A review on the long short-term memory model. *Artif. Intell. Rev.* **2020**, *53*, 5929–5955. [CrossRef]
44. Hossin, M.; Sulaiman, M.N. A review on evaluation metrics for data classification evaluations. *Int. J. Data Min. Knowl. Manag. Process* **2015**, *5*, 1.
45. Novaković, J.D.; Veljović, A.; Ilić, S.S.; Papić, Ž.; Milica, T. Evaluation of classification models in machine learning. *Theory Appl. Math. Comput. Sci.* **2017**, *7*, 39–46.
46. Orrù, P.F.; Zoccheddu, A.; Sassu, L.; Mattia, C.; Cozza, R.; Arena, S. Machine learning approach using MLP and SVM algorithms for the fault prediction of a centrifugal pump in the oil and gas industry. *Sustainability* **2020**, *12*, 4776. [CrossRef]