The Nucleus ***60***, *No. 2 (2023) 145-152*

https://doi.org/10.71330/thenucleus.2023.1306

www.thenucleuspak.org.pk

The Nucleus

ISSN 0029-5698 (Print)

ISSN 2306-6539 (Online)

A Comparison of Address Matching Techniques to Improve Geocoding: A Case Study of Islamabad, Pakistan

Abdul Basit Bashir, Muhammad Farooq Iqbal*

Applied Geo-Informatics Research Lab, Department of Meteorology, COMSATS University Islamabad (CUI), Park Road, Chak Shahzad, Islamabad, Pakistan

ABSTRACT

Geocoding is the process of converting addresses into spatial coordinates. It has become a need for the modern world in many fields such as strategy making, disaster management, location-based analysis, and planning of infrastructure, etc. Address matching is of critical importance in geocoding and is dependent on the language, address format, and components of the addresses. Different algorithms, tools, and applications have been designed for improving address matching. The goal of the research was to standardize geocode addresses using different Natural Language Processing (NPL) models. To standardize and geocode distinct types of addresses, Islamabad was chosen as the study area because it incorporates standard and unusual addresses. Address datasets were obtained from telecommunication industries operating in the study area. In this research, two NLP models including DeepParse (DP) and SpaCy's Named Entity Recognition (NER) were utilized and trained for address standardization. Each of the models operates using different techniques for parsing the addresses. The SpaCy model performed well with an accuracy of over 80% in all types of addresses. The DP model only outperformed the SpaCy model in urban areas with an accuracy of 95%, the others were less than 80%. This study focused on what types of addresses must be used to improve geocoding. Methods discussed in this study would be helpful in achieving a higher success rate for improving address matching percentage in the geocoding process. This research will help in choosing the best technique to improve address matching of addresses in Pakistan.

Keywords: *Geocoding, Address standardization, Address matching, Natural Language Processing, DeepParse, SpaCy, Named Entity Recognition*

1. Introduction

In this era of digitalization, development, and extensive applications of Geographic Information System (GIS), geocoding plays a significant role in filling the gap between non-spatial and spatial data [1]. Geocoding is the process of converting addresses into spatial coordinates. It has become a need for the modern world in many fields such as strategy making, disaster management, location-based analysis, and planning of infrastructure, etc., to be mentioned [2]. Address matching is of critical importance in geocoding, because of which geocoding is quite difficult in many rural or poorly mapped areas. Different algorithms, tools, and applications have been designed for improving address-matching percentages. Address matching is dependent on the language, address format, and the components of the address [3].

An address contains information about a specific place on a map. This information includes name, street, city, and province. Despite having address components, there is still a lack in positional accuracy of addresses. To improve this accuracy, geocoding comes in where addresses are assigned geographical coordinates i.e., latitude and longitude values [1]. Researchers use geocoding in various fields such as environmental variables leading to diseases [4], error corrections for InSAR point cloud [5], crime mapping [6], public safety, route directions [7], the relationship between household environment and health consequences [8], environmental epidemiology [9], locating traffic crashes [10] and the need of predictive policing [11]. Researchers used census data to geocode households and individual's addresses [12]. The addresses are collected on a daily, monthly and annual basis by many organizations such as educational institutions, health departments, police stations, courts etc. Many applications have been developed to geocode addresses, and much literature has been written on how to improve the match rates, and which technique should be used for improved accuracy at the local or national level [2].

Multiple models and algorithms have been designed for reducing spelling errors and improving the accuracy of geocoding of the street-level addresses. Levensthein et al introduced a method efficient enough to correct, delete, insert, and reverse each string to convert one string to another [13]. Perez-Diez et al used SpaCy's NER in the identification of medical texts written in Spanish from radiology reports and achieved a precision score of 99.87%, recall score of 99.28%, and F1 score of 99.58% [14]. In another study, the same model was used to identify names of the places from historical masses [15]. Another method introduced by Western Airlines in 1977 was a match rating computer approach focused on comparing and indexing homonymous words [16]. Repeat and near repeat analyses help in learning the sensitivity of repetition to geocoding [17]. Zandbergen et al [18] compared three data models including street network, parcel boundaries and address/point and it was suggested that address point data model is more effective for positional accuracy and reaching maximum match rate. Spatial point pattern test developed by Andersen can help to find similarities and differences in one or more spatial point patterns [19]. Geocoding was also used to solve classification difficulties, for this LGGeoCoder; a deep learning neural network was developed [3]. A study conducted by Abid et al [20] trained the deep learning model using optical character recognition (OCR) that takes input in an image form and returned the classified address, BLSTM layer was also used to improve the results. Study assessed

*Corresponding author:farooqbuzdar@gmail.com

the model on addresses of the United States of America, where usually addresses are not complex and 90.44% accuracy was achieved.

A lot of work on improving geocoding address matching has been done. In India, due to invalid address formats, learning algorithms were designed to understand and solve errors in addresses [21]. To deal with many unstructured address data, a deep learning-based address matching method was introduced which helped to identify semantic similarities between address records [22]. China faces the same issue regarding addresses, the major issue is the language barrier. Since the Chinese language has no delimiters among words, an optimized Chinese address-matching method was built to improve geocoding, address standardization, address modeling, and matching [2]. Dilek et.al. [23] designed a processing system through which communication with the device was made easier. To improve geocoding accuracy and address matching in Turkey, Matchi et al. [24] used two NLP techniques, i.e., Levenshtein distance algorithm (LDA) and match rating approach to standardize non-standard Turkish addresses, and concluded that geocoding can be used to track crime, monitoring of epidemics. Another method to update older tree inventories with coordinates by means of street-level panorama images and a global optimization framework for tree occurrence matching, geolocations of tree inventories were obtained using street addresses and a global positioning system (GPS) [25]. A study was conducted on spatial mobility patterns of victims and child abuse offenders under the crime mobility triangle methodology which is useful for criminal investigation in France [26].

Pakistan is the country where addresses are written in Roman Urdu and it is the first and foremost problem in address matching, as computers can neither read nor write Roman alphabet. Other problems include spelling mistakes, missing address components, and abbreviations. Addresses vary from simple to complex, there are developed cities that have standard addresses and cities with incomplete addresses. These issues lead to inaccurate geocoding and less match ratio. Thus, to improve the match ratio, addresses need to be standardized.

The core objective of this study is to improve geocoding by standardizing addresses using Natural Language Processing (NLP) techniques. The second objective is to develop an address parsing model that helps communicate with the computer in natural language. Lastly, to compare geocoded addresses using the address parsing models.

2. Data and Methodology

2.1 Study Area

Islamabad, the capital of Pakistan lies at 33.68°N and 73.04°E, at an elevation of 1475 meters with moderate climate [27]. The total area of the capital city is 906 km² and is divided into five zones including commercial, residential, and diplomatic areas. Out of these five zones, only zone-I and zone-II were developed following the master plan. Zone-I and zone-II are assigned for urban areas whereas the rest of the zones are assigned for rural areas. Urban zones are further divided into sectors and each sector comprises 2 km² with a central shopping mall [28]. Fig. 1 shows the study area map with zones and elevation.

2.2 Datasets

Address data used in this study was obtained from Pakistan telecommunication company limited (PTCL) because it operates both in rural and urban areas. PTCL provided data in comma-delimited format (.csv). After making sure that file contained only addresses of the study area, the address data was then divided into six categories including residential, urban, commercial, banks, rural, and schools. To train machine learning (ML) models specifically, DeepParse (DP) and SpaCy, another dataset was extracted via Google scrapping methods containing a total number of 349,637. Distinct types of addresses were then parsed through these trained models. To assess the accuracy of the trained models, ground data was collected using GPS by doing a ground survey of the addresses collected from PTCL. Random GPS points of the 100 addresses were taken to validate the results obtained from the models.

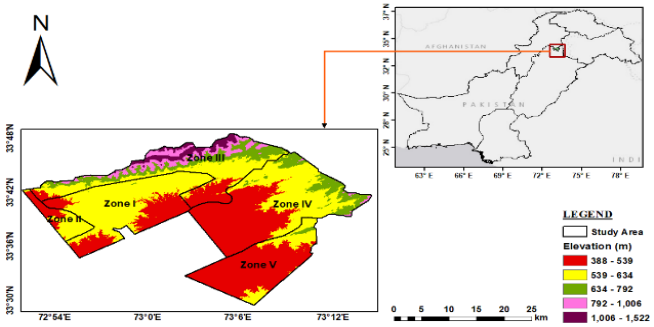

Fig. 1: Study area map of Islamabad showing zones along with the elevation in meters.

2.3 Methodology

2.3.1 Data Preparation

The ML models are typically predictive in nature and thus require preprocessed data to perform proficiently. The acquired dataset contains a lot of ambiguous data which decreases the model's accuracy and does not allow it to train the model. Therefore, data must be passed through the cleaning phase, where white spaces, illegal characters, duplicates, and repetitions of words in addresses, commas, and periods were all removed. After cleaning the data, it was prepared to make the model's prediction better. In this process, labels were assigned to each component of the address to make the model's prediction accurate.

2.3.2 Address Standardization Using Natural Language Processing Techniques

The NLP removes the communication barrier between computer and human language using artificial intelligence (AI) techniques [24]. The NLP-based techniques identified spelling errors and multiple spellings of any word in any

component of an address. A dictionary was created in which all the possible spellings and the correct spelling of words were listed and programmed in such a way that it displayed the correct addresses database and generator. Parser read dictionaries and broke the address into different bits. Analyzer read all the bits of the addresses, checked, and corrected errors using named entity recognition (NER) and DP model. These models checked for names, street numbers, and city names to produce accurate spellings for each address fragment. Output of an address, major components of NLP were the dictionary, parser, and semantic analyzer. Fig 2 demonstrates the methodology for the research.

2.3.2.1 SpaCy's Named Entity Recognition Model

The NER is one of the NLP methods to extract information from a text. The NER model for this study was trained using SpaCy which is an open-source NLP library that uses neural network or deep learning formulas to employ NLP models for embedding, encoding, attending, and predicting components in tasks such as parts of speech tagging, NER, and parsing etc. For this study, the model was trained in four steps (i) embedding, (ii) encoding, (iii) attention, and (iv) prediction.

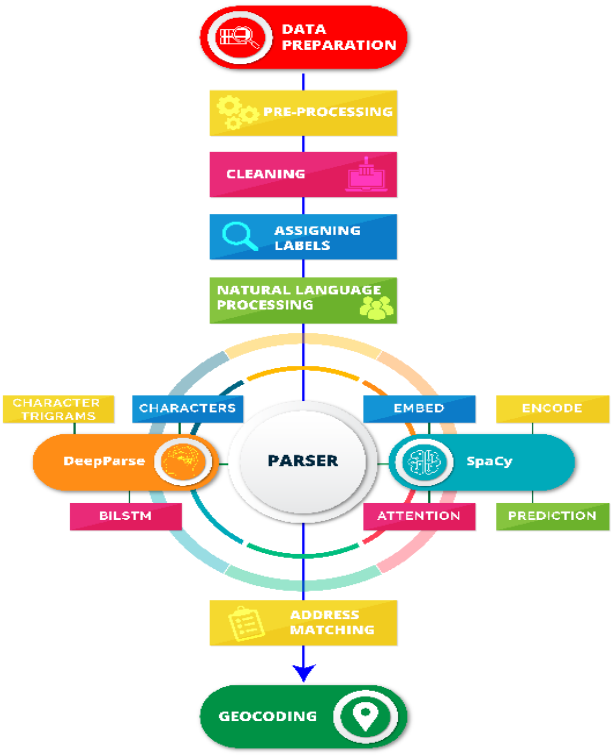

Fig. 2: Methodology of how the addresses were first standardized using NLP method.

The embedding stage of SpaCy uses prefixes, suffixes, shapes, and upper/lower case words to extract hashed values that reflect the similarity of words. In the encoding stage, hashed values passed through a convolutional neural network (CNN) so that values get fused with their framework. The result of the encoding stage to represent information is in the form of a matrix of vectors. Before the prediction of a tag, the matrix must pass through the attention layer of the CNN, this sums up the input and passes it onto the prediction stage. At the prediction stage, an entity ruler (ER) function was used to predict the tag for the available entity. The ER function has a variety of rules or patterns to allow the SpaCy model to predict and place each entity in the correct location. Since, ER function reads patterns of the words, and in the current study, addresses have unique patterns which the model can't predict or read. For the model to predict a tag for any entity in the given address, a new rule was designed to define the pattern of sectors, cities, and countries. This is how ER function was pre-trained before it could predict the location of any entity.

Table: 1 Multiple spelling of the same word to help the model read the patterns.

Tag	Check 1	Check 2	Check 3	Check 4
Islamabad	Isl	Isb	Isld	Islamabad
Pakistan	Pk	Pak	Pakistan	-
Sector	I8/1	G9/2	E8/3	F6/4

Table 1 shows every possible spelling of the city, country, and sector's patterns. Multiple spellings of the same word were first identified in the data, which was fed to ER to predict the right tag for every component of the address. Every sector and its sub-sectors were identified and passed to the ruler. This is how the model was trained.

For the model's evaluation precision, recall, and F_1 scores were calculated using Eq. (1) to (3).

$$P = \frac{C}{M} \tag{1}$$

$$R = \frac{C}{N} \tag{2}$$

$$F_1 = \frac{2PR}{P+R} \tag{3}$$

Where P is the precision, R is recall, C is the number of corrected entities, M is the number of matched entities with the reference data and N is the total number of predicted entities.

2.3.2.2 DeepParse Model

The DP model is a trainable postal address parser to parse global street addresses based on Recurrent Neural Network (RNN) architecture that uses Long-Short Term Memory (LSTM) Bi-directional Long-Short Term Memory (BLSTM) to parse non-standardized addresses [20]. Since postal addresses are memory dependent and cope with sequential issues.

Hochreiter et al [29] introduced LSTMs to cater to information that must be remembered throughout the model training process. The DP model works in three steps including (i) characters, (ii) character trigrams, and (iii) words. It takes the address as input and divides each word into a list of words. After making a list of words, each word was then divided into an alphabet/character and then trigrams were made. These trigrams were then embedded to convert into vectors to make these words pass through RNN

layers that put and classify the words to their respective attribute.

Fig. 3 shows the workflow of the DP model. The DP model takes input in three formats including words, characters, and character trigrams. These inputs were used to generate vector representation and pass through the embedding stage. After getting the vectors of each word, these were then passed on to BLST layer individually and studied their picture. The dropout layer of BLSTM was then applied to each word and it ignores irrelevant information, thus producing only valuable information. The projection layer of BLSTM makes the output data appropriate according to the dimensions, by taking the output of BLSTM and the dropout layer, linking them, and making the data ready for another layer i.e., the learning layer. The learning layer concatenates the projected layer's output and embedded words. It works on the classification by studying the features at the word level making the model learn patterns effectively and classifies it accordingly [22]. To train the model fasttext (FT) library of words for text learning and classification was used. The FT helps in processing the words before any parsing model starts working. The FT is a quick and efficient method that learns and represents words in vector form and classifies it [30]. The FT instead of learning the representation of words, it focuses on the core structure of words. This is important in the interpretation of different morphological forms of words separately [31].

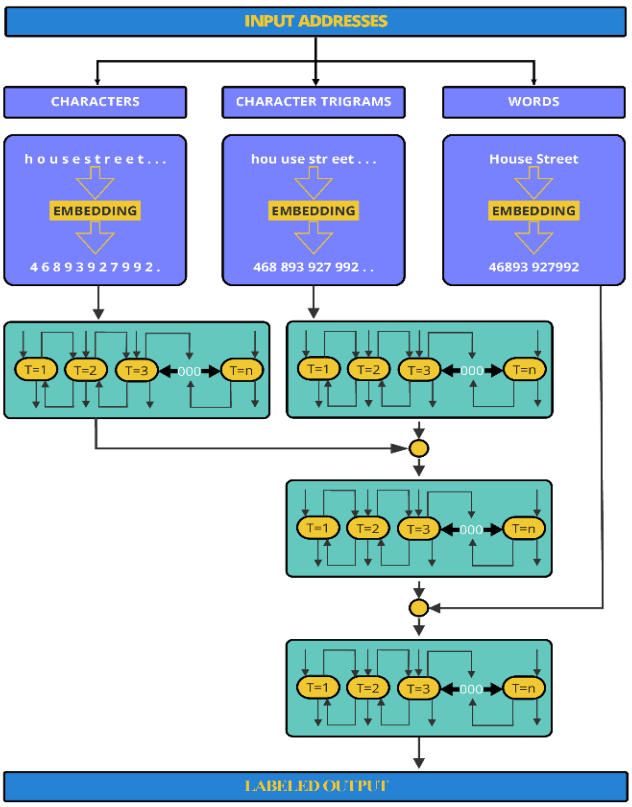

Fig. 3: The workflow of the DP model including the embedding of characters, characters trigrams, and words-for-word predictions.

This model was trained on 26,000 textual addresses of Islamabad. These addresses were divided into two parts, 80% of the addresses were used to train the model whereas the remaining 20% was used for testing the model. After completing all the pre-processing steps, the model training started by setting epoch to 5, batch size to 8, and learning rate to 0.01. A learning rate scheduler was also implemented to reduce the learning rate per epoch, this prevented the model from overtraining or overfitting. After training DP model, its FT function was again retrained for the model's fine-tuning and another addition of an attention layer that pays close attention to every single address before delivering the output. This is how the model was trained.

Lastly, to predict the model's results, a CSV file in which preprocessed, cleaned, and labeled addresses of urban, rural, residential, and commercial, banks and schools were imported. Moreover, pandas, urlopen, and json libraries were imported. A loop from 0 to the end of the column was defined, where each address in the files was passed to the model and returned the formatted or parsed address.

The matched and unmatched addresses were checked by passing both original and parsed addresses using Google API. This was done by comparing the latitude of the original address with the latitude of the parsed address, and their match score was achieved. These parsed addresses were then passed on to the Google API that returned the latitude and longitude of the addresses and were geocoded.

2.3.3 Statistical Analysis

2.3.3.1 Spatial Point Pattern Test

A spatial point pattern test developed by [32] was used to identify the similarities and differences among points. The test's initial use involved a comparison of the spatial patterns of various crime types. Their values ranged between 0 and 1. Where 0 represents no similarity and 1 represents ideal similarity. This test was applied in numerous situations. The test was also applied to the following research topics: altering international trade patterns [33] and crime's spatial seasonality [34].

Using the results of the above-mentioned statistical tests, a similarity index between the results of the three datasets was computed. The similarity index, S was used to assess how similar the two datasets are to one another. The value ranged between 0 and 1, where 0 represents no similarity, while 1 is the perfect similarity. Since the values were too small to find the similarities and differences, tolerance was also set to 0.003m. A similarity index was calculated using Eq. (4).

$$S = \frac{\sum_{i=1}^{n} Si}{n} \tag{4}$$

Where, S is the similarity index and S_i is 1, if the point pattern of two datasets is the same and n is the number of points taken.

2.3.3.2 Mean Absolute Error and Root Mean Square Error Metrics

The mean absolute error (MAE) and root mean square error (RMSE) statistical metrics are employed frequently in model evaluation studies. To calculate RMSE, the distance between the points i.e., from ground points to SpaCy's and from ground points to DP were calculated using Eq. (5). Once distance was computed, it was used to calculate MAE and RMSE of both the datasets i.e., SpaCy and DP, considering ground data to be accurate. Eq. (6) and (7) were used to calculate MAE and RMSE respectively.

$$d\,(g, m) = \sqrt{\sum_{i=1}^{n}(g_i - m_i)^2} \quad (5)$$

Where *d* is the distance, *g*, and *m* are the points, g_i, and m_i are Euclidean vectors in space, and *n* is space.

$$MAE = \sum_{n=1}^{i}|g - m| \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{n=1}^{n}(g - m)^2} \quad (7)$$

Where *n* is the total number of observations *g* is the ground data and *m* is the data obtained from models.

3. Results

3.1 Address Standardization Using NLP

3.1.1 DeepParse Model

Fig. 4a shows the geocoding results of the addresses parsed by DP model. The DP model performed well in urban areas of Islamabad with a 95% accuracy rate. A total of 80 addresses were parsed and geocoded out of which only 4 were not matched, the remaining 76 were perfectly matched and geocoded. Rural areas of Islamabad in Zone-III have complex addresses with missing components and addresses were written in Roman which couldn't be parsed well. Out of 95 addresses for rural 48 were parsed and matched whereas 47 addresses remained unmatched. This mismatch affected geocoding accuracy which dropped to 50.53%. Because of address components such as flat/apartment number, floor number, and house names instead of house numbers in the address data for residential areas, such addresses couldn't be parsed correctly and didn't find any matching address in the Google geocoding API.

A total of 43 residential addresses were parsed by DP model where 18 addresses were matched, and 25 addresses were unmatched with an accuracy of 41.86% geocoding. Commercial areas usually have simple addresses and are easy to match and geocode, but this accuracy drops in some cases where addresses include shop number, floor number, or a building/plaza number. Such information couldn't be parsed and geocoded well. However, geocoding results for the commercial areas of Islamabad were 83.72%. For commercial addresses out of 43 addresses, 36 were matched and 7 were unmatched.

Banks are located in both urban and rural areas. Addresses are easy to parse in urban areas but in some cases plots, shops, or building numbers have affected the results whereas, in rural areas, khasra numbers in the addresses were

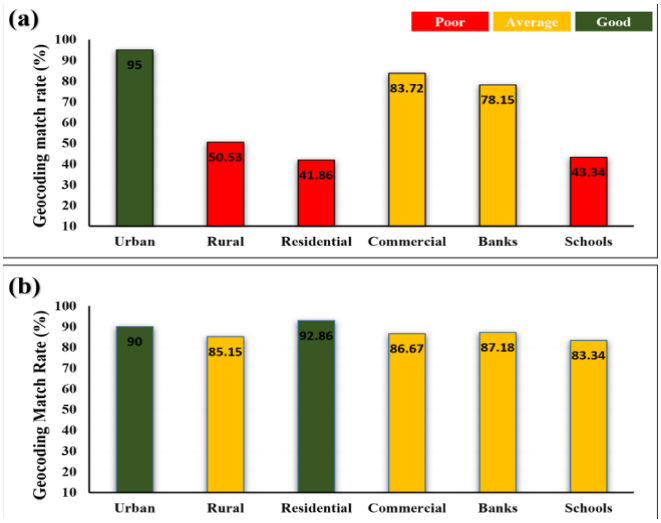

Fig. 4 (a) Number of geocoded addresses in percentage with Google API by using standardized addresses of DeepParse model and (b) displays number of geocoded addresses in percentage Google API by using standardized addresses of SpaCy model.

hard to parse by the model. A total of 41 bank addresses were parsed, out of which 32 addresses matched and 9 addresses were unmatched. The total geocoding accuracy for the banks was 78.15%. For schools, a total of 30 addresses were parsed out of which only 13 were matched and 17 were unmatched, making geocoding accuracy of 43.34%. This is because schools don't have proper addresses, they are usually searched by their name, and one school may have multiple branches and all branches can't have the same latitude and longitude values, hence reducing geocoding results. Fig. 5a shows the comparison for matched and unmatched addresses of all types of addresses using DP model.

3.1.2 SpaCy's Named Entity Recognition Model

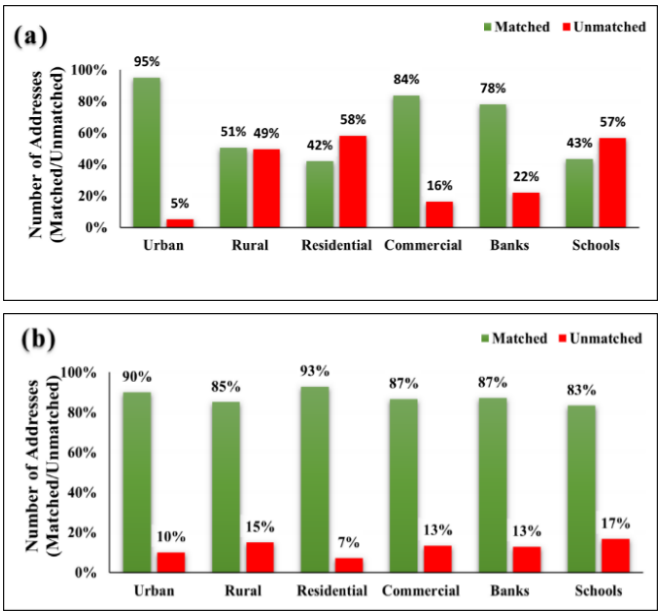

Fig. 5 (a) Comparison of matched and unmatched addresses with Google Geocoding API using DeepParse model and (b) Comparison of matched and unmatched addresses with Google Geocoding API using SpaCy model.

Fig. 4b shows the geocoding results for SpaCy's NER model that were performed in all types of addresses including urban, rural, residential, commercial, banks and schools. A total of 80 urban addresses were parsed out of which 72 were correctly parsed and matched whereas the remaining 8 addresses didn't parse and remained unmatched with an accuracy of 90% geocoded addresses. Out of 87 addresses of rural areas of Islamabad, 74 addresses were parsed and matched with an accuracy of 85.15% geocoded, while 13 didn't parse and remained unmatched. A total of 42 addresses for residential areas were parsed, of which 39 were parsed and matched, whereas 3 didn't parse and matched, thus making geocoding accuracy of 92.86%. Geocoding accuracy for commercial addresses was 86.67% where 39 addresses out of 45 were perfectly matched. Around 87.18% and 83.34% of addresses were parsed, matched, and geocoded for banks and schools, respectively. Fig. 5b shows the comparison of matched and unmatched addresses of urban, rural, residential, commercial, banks, and schools using the SpaCy model.

3.2 Statistical Analysis

3.2.1 Spatial Point Pattern Test

The comparison of both the DP and SpaCy models showed that the SpaCy model performed better in all types of addresses with an accuracy of more than 80%, while DP model showed variations depending on the region or type of addresses. Fig. 6 shows the statistical comparison of both models. In urban areas, the DP model gave an accuracy of 95% while the SpaCy model had an accuracy of 90%. In rural areas, SpaCy model outperformed DP model with an accuracy of 85.15%, whereas DP model could only achieve 50.53%. In residential areas, an accuracy of 41.86% and 92.86% were achieved by DP and SpaCy models respectively. Both models performed well in commercial areas with a difference of 2.95%. The DP model observed 83.72% accuracy whereas the SpaCy model observed 86.67% accuracy. Around 78.14% and 87.18% accuracies were achieved by DP and SpaCy models respectively for the addresses of banks. The SpaCy model hit an accuracy of 83.34% for school addresses, whereas the DP model resulted in an accuracy of 43.34%.

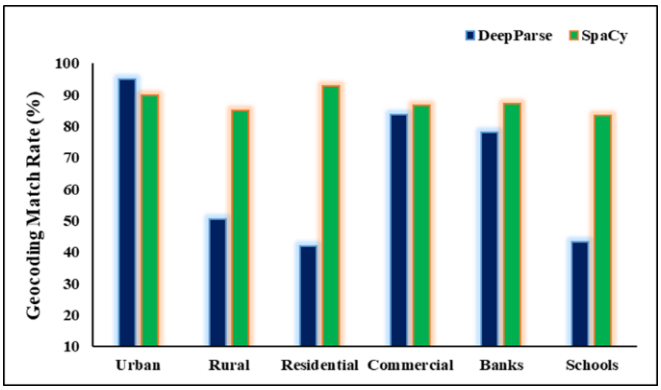

Fig. 6: Comparison of geocoded results achieved by using DeepParse and SpaCy models.

3.2.2 MAE and RMSE Metrices

The MAE of the SpaCy model was 0.10 m whereas 0.12 m was observed for the DP model. The RMSE for the SpaCy model was 0.26 m whereas for DP model it was 0.32 m. The MAE and RMSE of both the models show that the SpaCy model turned out to be more accurate as it had less MAE and RMSE, when addresses were compared with the ground data. Table 2 shows the distance between the GPS ground points and the points obtained from DP and SpaCy models. Fig. 7a illustrates the classified distance between the DP model's geocoded points and GPS points while 7b illustrates the classified distance between the SpaCy model's geocoded points and GPS points.

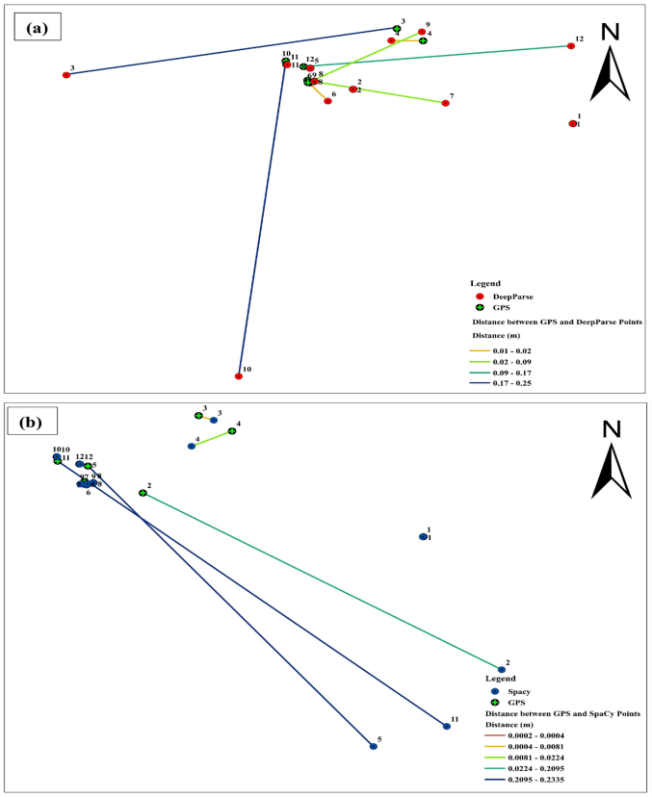

Fig. 7: (a) Classifications of distance (m) between GPS and DP model points, and (b) classifications of distance (m) between GPS and SpaCy model points.

4. Discussions

Two address parsers including DP and SpaCy models were trained to read, parse, and geocode addresses. In the current study, DP model was trained using the FT function of BLSTM on different types of addresses in Islamabad with addresses from simple to complex. The overall accuracy calculated was 65.43%. This is mainly due to the complex nature of addresses, as these are written in Roman, and in some areas, these don't have complete address components, thus making it difficult to be parsed by the parser. The model is best performed in urban areas with 95% accuracy rate as these follow a standard address pattern.

The second model used in this research was SpaCy's NER, and it turned out to be a good parser because of its

ability to understand and learn the patterns. This was trained on 26,000 Islamabad addresses to calculate the precision, recall, and F1 score. Results obtained by NER were totally dependent on the training data, it can't check the sequence of the words like DP, and it depends on the patterns provided for the training. NER techniques were applied to our dataset and a precision score of 93.90%, recall score of 92.86%, and F1 score of 93% was achieved.

The results obtained from the models helped in standardizing the address to be geocoded more accurately.

These results can help the courier services to find the address efficiently and fast, it can also help the food delivery services to reach the exact location instead of asking customers for the exact location.

Future researchers can use these techniques to study the addresses and geocode them in their own areas of interest. Geocoding can be performed to study the crime rate in any area or to control traffic accidents at any spot. Every city has different checks in their addresses, we could only address some of the checks in Islamabad, and these models can be trained by using addresses throughout Pakistan.

5. Conclusions

In this study, two address parsing models were utilized and trained to improve geocoding results in the capital city of Islamabad. The DP and SpaCy's NER models were used to parse addresses, their accuracy was accessed, and their results were compared. The DP model which is an international address parser could only perform well in urban areas with 95% accuracy rate. Whereas, SpaCy model outperformed in all types of addresses with an overall accuracy of 80% from assessing simple to complex addresses. The address data is written in both English and Roman, so it is quite difficult for the computer to read and parse the addresses.

For accurate address matching and geocoding, all the addresses were standardized using NLP techniques. The NLP methods identified and removed duplicate words, spelling errors, abbreviations, and ambiguous data. The techniques used in this study also removed the communication barrier between human and computer language, thus making address standardization easy. Two NLP models including DP and SpaCy's NER were picked and trained for address geocoding. Each of the models operates using different techniques for parsing the addresses. The DP model uses ML approaches that include bidirectional long short-term memory. This method memorizes the training data and presents results based on the training data provided. Whereas SpaCy model reads the patterns of the words, embeds, encodes and passes these addresses through the attention layer of the convolution neural network layer and then predicts the output. The results of both models were then compared. Both models performed well in urban areas with an accuracy of over 90% each. This was achieved because urban areas in the capital city are planned and follow standard address patterns. The SpaCy model outperformed DP model in rural, residential, commercial, bank and school address data. Because of its ability to read the patterns, it could easily read and predict the address word by word. It was trained in patterns, this is why it could easily read house, street, sector, and city information, even if these addresses were abbreviated.

In conclusion, SpaCy can be an efficient and effective solution for addressing standardization because of its ability to read patterns using NLP techniques. SpaCy's model efficiency and performance have been demonstrated in this study, thus making it a valued model to manage large datasets.

Acknowledgment

The authors would like to express their sincere gratitude to Pakistan Telecommunication Company Limited for giving the address data. Their cooperation in sharing this dataset has greatly helped the success of the study, and genuinely appreciate their support.

Table 2: The distance between the ground points acquired using GPS and the points obtained from DP and SpaCy models.

SpaCy			GPS		DeepParse		
Latitude	Longitude	Dist (m)	Latitude	Longitude	Latitude	Longitude	Dist (m)
33.657	73.157	0.0002	33.111	73.157	33.652	73.157	0.0184
33.570	73.196	0.0004	33.679	73.016	33.679	73.016	0.0203
33.722	73.052	0.0081	33.725	73.044	33.689	72.832	0.2146
33.714	73.040	0.0225	33.716	73.061	33.716	73.040	0.1721
33.525	73.132	0.2095	33.695	72.988	33.695	72.988	0.0904
33.683	72.988	0.2224	33.683	72.988	33.669	73.000	0.0830
33.684	72.985	0.2345	33.685	72.987	33.668	73.075	0.2461
MAE	0.10					0.12	
RMSE	0.26					0.32	

References

[1] F. Melo and B. Martins, "Automated geocoding of textual documents: A survey of current approaches," Transactions in GIS, vol. 27, no. 2, pp. 3-38, 2017.

[2] Q. Tian, F. Ren, T. Hu, J. Liu, R. Li, and Q. Du, "Using an optimized Chinese address matching method to develop a geocoding service: A case study of Shenzhen, China," ISPRS International Journal of Geo-Information, vol. 5, no. 11, pp. 650-667, 2016.

[3] Z. Yan, C. Yang, L. Hu, J. Zhao, L. Jiang, and J. Gong, "The integration of linguistic and geospatial features using global context embedding for automated text geocoding," International Journal of Geo-Information, vol. 10, no. 9, pp. 572, 2021.

[4] S. E. K. M. Rompa et al., "An approach to the asthma-protective farm effect by geocoding: good farms and better farms," Pediatric Allergy Immunol, vol. 29, no. 3, pp. 275-282, 2018.

[5] S. Montazeri, F. R. Gonzalez, and X. X. Zhu, "Geocoding error correction for InSAR point clouds," Remote Sensing, vol. 10, no. 10, pp. 1523, 2018.

[6] C. White and D. Weisburd, "A co-responder model for policing mental health problems at crime hot spots: Findings from a pilot project," Policing: A Journal of Policy and Practice, vol. 12, no. 2, pp. 194-209, 2018.

[7] K. J. Sydow, S. Stobernack, and S. Wienecke, "Accuracy of address geocoding with GIS: A data analysis of households in a German City," Transactions in GIS, vol. 12, no. 3, pp. 333-354, 2008.

[8] B. Wilson, N. Wilson, and S. Martin, "Using GIS to advance social economics research: geocoding, aggregation, and spatial thinking," Forum for Social Economics, vol. 50, no. 4, pp. 480-504, 2021.

[9] E. J. Kinnee, S. Tripathy, L. Schinasi, J. L. C. Shmool, P. E. Sheffield, F. Hpuguin, and J. E. Clougherty, "Geocoding error, implications for exposure assessment and environmental epidemiology," International Journal of Environmental Research and Public Health, vol. 17, no. 16, pp. 5845, 2020.

[10] X. Qin, S. Parker, Y. Liu, J. A. Graettinger, and S. Forde, "Intelligent geocoding system to locate traffic crashes," Accident Analysis and Prevention, vol. 50, pp. 1034-1041, 2012.

[11] J. M. Caplan, L. W. Kennedy, E. L. Piza, and J. D. Barnum, "Using vulnerability and exposure to improve robbery prediction and target area selection," Applied Spatial Analysis and Policy, vol. 12, no. 1, pp. 113-136, 2019.

[12] N.S. Walford, "Bringing historical British population census records into the 21st century: A method for geocoding households and individuals at their early-20th century addresses," Population Space and Place, doi: 10.1002/psp.2227.

[13] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," Soviet Physics – Doklady, vol. 10, pp. 707–710, 1965.

[14] D. Perez-Diez, D. Rodriguez, V. Blanco-Valero, A. Caceres, and G. Castellanos-Dominguez, "Identification of medical texts written in Spanish from radiology report using SpaCy'NER," Journal of Biomedical Informatics, vol. 113, pp. 103647, 2021.

[15] J. Won, J. Seo, J. Lee, and Y. Choi, "Named entity recognition of historical places in Korean Buddhist scriptures," Digital Scholarship in the Humanities, vol. 33, no. 2, pp. 401-412.

[16] G. B. Moore, "Accessing individual records from personal data files using nonunique identifiers. Final report. Computer science & technology series," 1977.

[17] C.P. Haberman, D. Hatten, J.G. Carter, and E.L. Pizza, "The sensitivity of repeat and near repeat analysis to geocoding algorithms," Journal of Criminal Justice, vol. 71, pp. 101721, 2021.

[18] P.A. Zandbergen, "A comparison of address point, parcel and street geocoding techniques," Computers, Environment and Urban Systems, vol. 32, pp. 214-232, 2007.

[19] M. Andresen, T. W. Jørgensen, and P. J. Diggle, "Spatial point pattern analysis of crime incidents: An overview and some open research problems," Statistical Science, vol. 35, no. 2, pp. 283-298, 2020.

[20] N. Abid, A. Hasan, and F. Shafait, "DeepParse: A trainable postal address parser," in Digital Image Computing: Techniques and Applications, pp. 1–8, 2018.

[21] V. Srivastava, P. Tejaswin, L. Dhakad, M. Kumar, and A. Dani, "A geocoding framework powered by delivery data," in Proceedings of the 28th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2020, pp. 1-4, doi: 10.1145/3397536.3422254.

[22] Y. Lin, M. Kang, Y. Wu, Q. Du, and T. Liu, "A deep learning architecture for semantic address matching," International Journal of Geographical Information Science, vol. 34, no. 3, pp. 559-576, 2020.

[23] A. E. Dilek, C. Yan, H. Jin and M. Bektasoglu, "An optimized Chinese address matching method based on address standardization, modeling and matching," Journal of Cleaner Production, vol. 171, pp. 1213–1223, 2018.

[24] D.K. Matchi and U. Avdan "Address standardization using the natural language process for improving geocoding results" Computers, Environment and Urban Systems vol.70, pp. 1–8, 2018.

[25] D. Laumer, N. Lang, N. Van Doorn, O. M. Aodha, P. Perona, and J. D. Wegner, "Geocoding of trees from street addresses and street level images," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 162, pp. 125-136, 2020.

[26] J. Chopin and S. Caneppele, "Geocoding child sexual abuse: An explorative analysis on journey to crime and to victimization from French police data," Child Abuse and Neglect, vol. 91, pp. 116-130, 2019.

[27] A. Aslam, I. A. Rana, and S. S. Bhatti, "The spatiotemporal dynamics of urbanisation and local climate: A case study of Islamabad, Pakistan," Environmental Impact Assessment Review, vol. 91, pp. 1-10, 2021.

[28] M.J. Butt, A. Waqas, M.F. Iqbal, G. Muhammad and M.A.K. Lodhi "Assessment of Urban Sprawl of Islamabad Metropolitan Area Using Multi-Sensor and Multi-Temporal Satellite Data" Arabian Journal for Science and Engineering, vol. 37:pp. 101–114, 2012.

[29] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[30] I. Santos, N. Nedjah, and L. M. Mourelle, "Sentiment analysis using convolutional neural network with fastText embeddings," in IEEE Latin American Conference on Computational Intelligence, 2017, doi: 10.1109/LA-CCI.2017.8285683.

[31] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Transactions of the Association for Computational Linguistics, vol. 5, pp. 135-146, 2017.

[32] M.A. Andresen, "Testing for spatial dependence in the distribution of property crime using point pattern analysis and GIS," Journal of Quantitative Criminology, vol. 25, no. 3, pp. 339-366, 2009.

[33] M. A. Andresen, "Canada-United States interregional trade: quasi-points and spatial change," The Canadian Geographer, vol. 54, no. 2, pp. 139-157, 2010.

[34] M.A. Andresen and N. Malleson, "Spatial heterogeneity in crime analysis," in Crime Modeling and Mapping Using Geospatial Technologies, M. Leitner, Ed. New York, NY: Springer, pp. 3-23, 2013.