Journal of Geographical Systems
https://doi.org/10.1007/s10109-023-00435-8

ORIGINAL ARTICLE

Unveiling the impact of machine learning algorithms on the quality of online geocoding services: a case study using COVID-19 data

Batuhan Kilic[1] · Onur Can Bayrak[1] · Fatih Gülgen[1] · Mert Gurturk[1] · Perihan Abay[2]

Received: 11 January 2023 / Accepted: 17 November 2023

Abstract
In today's era, the address plays a crucial role as one of the key components that enable mobility in daily life. Address data are used by global map platforms and location-based services to pinpoint a geographically referenced location. Geocoding provided by online platforms is useful in the spatial tracking of reported cases and controls in the spatial analysis of infectious illnesses such as COVID-19. The first and most critical phase in the geocoding process is address matching. However, due to typographical errors, variations in abbreviations used, and incomplete or malformed addresses, the matching can seldom be performed with 100% accuracy. The purpose of this research is to examine the capabilities of machine learning classifiers that can be used to measure the consistency of address matching results produced by online geocoding services and to identify the best performing classifier. The performance of the seven machine learning classifiers was compared using several text similarity measures, which assess the match scores between the input address data and the services' output. The data utilized in the testing came from four distinct online geocoding services applied to 925 addresses in Türkiye. The findings from this study revealed that the Random Forest machine learning classifier was the most accurate in the address matching procedure. While the results of this study hold true for similar datasets in Türkiye, additional research is required to determine whether they apply to data in other countries.

Keywords Address matching · COVID-19 · Geocoding · Machine learning · Random forest

JEL Classification C45 · C52 · C53 · I18

Extended author information available on the last page of the article

Published online: 25 January 2024

1 Introduction

Building a link between a georeferenced textual description and its appropriate coordinate pair on the surface of the Earth is called geocoding (Zandbergen 2008). It involves comparing the input data with a benchmark dataset, identifying the candidate matches, and returning the standardized address and/or place name along with its geographic coordinates. Geocoding is typically performed within a Geographical Information System (GIS) environment, either offline or online. Conventional geocoding is carried out using the tools included in GIS software packages under the supervision of GIS professionals. Developments in Internet technology and the proliferation of commercial companies, including ArcGIS Online, Bing Maps, Google Maps, and HERE Maps, have boosted the use of online geocoding processes. The geocoding method has been widely employed in different fields of real-world applications such as public health and epidemiology (Rushton et al. 2006; Goldberg and Cockburn 2012; Goldberg et al. 2013), or safety and crime analysis (Levine and Kim 1998; Bichler and Balchak 2007; Ratcliffe 2004; Qin et al. 2013). Recently, it has been utilized to implement the geographical tracking and control operations of contagious and lethal disease cases such as COVID-19 that have emerged worldwide with precision (Akakba and Lahmar 2020; Karabegovic et al. 2021; Cohen et al. 2022; Kilic et al. 2022). Over the past fifteen years, numerous studies have been conducted to address and evaluate geocoding methods from different perspectives (Davis and Fonseca 2007; Zandbergen 2009; Roongpiboonsopit and Karimi 2010; Cui 2013; Khalifa et al. 2017; Kilic and Gülgen 2020a).

While geocoding is efficiently utilized in numerous location-based research fields, the accuracy of the procedure has a direct impact on the quality of these studies. Geocoding accuracy is commonly evaluated using terms such as match rate and positional accuracy (Zandbergen 2008). The match rate, also referred to as completeness, is the percentage of matched records between the input data and the geocoder's standard dataset. Ratcliffe (2004) suggests that a match rate of approximately 85 percent is necessary to establish a statistically reliable model. Nonetheless, there are some studies where this rate has been described as lower or higher in different regional studies (Andresen et al. 2020; Briz-Redón et al. 2020). At the core of geocoding is the address, which serves as a syntactic description of a specific geographic location. However, various issues, such as missing or misspelled addresses and variations in the use of abbreviations, diminish the quality of geocoding results. Moreover, the reference address dataset itself may contain syntax errors that degrade geocoding accuracy. The positional accuracy of geocoding is determined by the Euclidean distance between the geocoded and the actual location. This distance indicates how close the geocoded point is to the "real" address location (Zandbergen 2009). Kounadi et al. (2013) noted that the geocoded position is accurate if the distance is less than 100 m. However, this threshold may vary depending on the characteristics of the specific study area. To assess the reliability of address matching, it is crucial to determine the level of similarity between an address and its corresponding entry in the reference database.

Text similarity measures are commonly employed to identify similarities between textual datasets, including address data. By searching the reference address dataset, they choose the standard address with the highest score that best matches the supplied data. In recent years, several text similarity algorithms focusing on geocoding techniques have been utilized (Koumarelas et al. 2018; Lee et al. 2020; Kilic et al. 2022). However, each of these metrics employs a unique calculation technique in the background, which makes it challenging to determine the most suitable metric for a specific dataset. Incorporating multiple text similarity algorithms into the process is a valid approach for analyzing the outcomes of address matching. The similarity scores generated by these algorithms play a crucial role in determining the accuracy of the matching process.

Machine learning, a subfield of artificial intelligence, originated from computational learning theory within the realm of computer science (Simon and Singh 2015). Machine learning algorithms enable decision-making systems through the learning and construction of prediction models on datasets (Sah 2020). These methods find extensive application in various research domains, such as biomedical image processing, web content filtering, recommendation systems, speech recognition, object detection in images, etc. (LeCun et al. 2015). In the field of geocoding services, machine learning-based classifiers can incorporate additional statistical predictors accessible through text similarity measures. Unlike traditional geocoding models, machine learning-based methods can increase the granularity of data analysis, category distinctions, and associated location information, which is particularly important and essential for address verification. Lee et al. (2020), depending on the quality of their simulated database, proposed a method to enhance the performance of the address matching process by integrating multiple similarity measures using machine learning. They have improved the accuracy of address matching to over 97% by employing nine text similarity metrics for three different machine learning models they trained. Existing research, however, does not focus on real-world applications, and there is a need to investigate extensively used online geocoding services, particularly in study areas where there is no geocoding standard. Therefore, our main objective is to address the following question: "Can machine learning classifiers present the reliability of address data retrieved from widely used online geocoding services?" In the direction of this research question, this study intends to reveal the capabilities of machine learning classifiers that can be used to measure the precision of address matching results produced by ArcGIS Online, Google Maps, Bing Maps, and HERE Maps geocoding services. As input data, addresses with detected cases of COVID-19 were utilized. To examine the address matching, the authors manually establish the binary relationship between each reference and the geocoding services' returned addresses. The addresses obtained from each geocoding service were compared with the benchmark data and categorized into two classes, true and false. If the retrieved address matches the reference address exactly, the geocoding result is encoded as true. In contrast, the dependent variable is categorized as false if any component of the address differs from the reference. The dependent variables of machine learning classifiers were defined as these classes. Seventeen text similarity metrics were utilized to generate input features. These metrics were employed to differentiate address accuracy and determine matched and non-matched addresses, enabling the extraction of address features. The objective of this research effort was to use machine learning to estimate address accuracy based on text similarity scores between the geocoding service and the reference address. In order to assess the impact of features on the classification models, various types of text similarity metrics, as well as their combinations, were experimentally included in the training and testing processes. Subsequently, the performance of the models was compared in terms of area under curve (AUC) scores and input feature groups in the final step.

The remainder of this paper is organized in the following manner. The second section presents a concise summary of related geocoding investigations and the Turkish addressing system. The next section gives an overview of the field of study and the methods used to measure how well machine learning classifiers improve the quality of online geocoding services. The third section provides the results and discussions and limitations of the study. Finally, the conclusion of this research is presented in the last section.

2 Related works on geocoding worldwide and in Türkiye

Geocoding has evolved into a standard practice within various research inquires, serving as a fundamental tool for conducting geographical analyses. However, most current geocoding systems still demonstrate notable spatial inaccuracies, primarily stemming from the inherent limitations of traditional geocoding methods and the constrained availability and dependability of the reference datasets used. The occurrence of spatial inaccuracies in the geocoding output has a consequential impact on the subsequent utilization of the resulting data in research investigations (Yin et al. 2019). In recent times, several studies have emerged with a focus on improving this process, particularly through the integration of artificial intelligence. Rashidian et al. (2018) devised an integrated geocoding model based on machine learning, which harnesses multiple complementary reference sources to attain high accuracy and mitigate address matching errors. Comber and Arribas-Bel (2019) utilized a conditional random field (CRF) token on address records, employed word2vec to convert them into vectors, and implemented three machine learning classifiers to make predictions based on a local database. Lin et al. (2020) introduced an address matching method based on deep learning to identify the semantic similarity between address records. They performed a comparative analysis with the results of different address matching methods and achieved 97% of accuracy address matching.

Geocoding is a globally difficult process, especially in Türkiye compared to the majority of Western nations. This is primarily due to the perpetually evolving addressing system in Türkiye, which makes obtaining accurate geocoding results uncertain and challenging (Matci and Avdan 2018; Kilic and Gülgen 2020b). The Turkish addressing system is comprised of multiple components, including street type or avenue, neighborhood, door number, postal code, district, and province (ARS 2007), with established abbreviations (PTT 2013). In Türkiye, studies on the standardization of the addressing system continue at a rapid pace, but there are numerous issues with the use of the system's components. Yildirim et al. (2014) have investigated the origins of mistakes in geocoded locations that were incompatible. They revealed that inaccuracies in external door numbers (35%) and street names (28%) led to mismatched addresses. In addition, various issues such as missing addresses (15%), misspellings (9%), missing offset data (5%), typos (4%), and inappropriate formatting (4%) also impact geocoding quality. Matci and Avdan (2018) identified various issues with the Turkish postal address due to typographical errors, misspellings, and non-standardized formats. Kilic and Gülgen (2020a) observed that the address similarity values produced from the two online geocoding methods for the Fatih district in Türkiye and the Miami Beach region in the United States differ by around 30%. Kilic et al. (2022) found that, on average, five distinct online geocoding services identified similarity rates of 57.3, 61.8, and 44.8% for descriptive address components such as neighborhood name, road name, and numbering.

A review of the relevant literature reveals that standardization issues negatively impact geocoding accuracy in Türkiye. Geocoding based on inconsistent data with complex address structures needs a robust accuracy analysis. In order to comprehensively address standardization issues related to address descriptions, several similarity metrics need to be examined. Incorporating these metrics into a decision support mechanism, especially a classification system based on machine learning, would facilitate the evaluation of the address matching process.

Broadly speaking, online geocoding services provide a limited degree of open-source functionality and do not provide users with direct access to or the ability to modify their own reference database. Furthermore, the specific text similarity approach employed for matching address elements remains unknown from end-users. To enhance the address matching process, the integration of machine learning-based approaches into global online geocoding research and applications, along with the incorporation of combined text similarity metrics, would effectively solve intrinsic limitations. While there have been studies aimed at improving the accuracy of address matching and enhancing the output quality of the geocoding process, there has been limited attention given to commercial systems due to the few detailed implementation descriptions. To our knowledge, to date, no work has studied the performance improvement of online geocoding services in detail. Also, there is a scarcity of research that investigates diverse text similarity metrics.

This study proposes a novel approach based on machine learning for solving the problem of reliable and accurate geocoding in different online geocoding services. In the implementation of geographical tracking and control processes for infectious and lethal disease cases, the performance of geocoding services in address matching assessment (matching and non-matching) is being investigated. For evaluating the address matching process, different machine learning methods with multiple text similarity metrics are employed. Additionally, the proposed approach utilizes hyperparameter tuning of machine learning methods in order to identify the best matching configuration. The training data for machine learning methods is sourced from residential addresses of individuals with documented cases of reportable infectious diseases. Concerning the quality of the training data, we also consider both the ratios of the matching and non-matching addresses of online geocoding services and different combinations of text similarity metrics to reveal the effectiveness of machine learning techniques.

3 Methodology

3.1 Study area and dataset

In accordance with statistics provided by the Turkish Ministry of Health, Istanbul, which accounts for about 40% of patient cases in Türkiye, has been selected as the study region for the investigation of test procedures (URL-1). More than 16 million instances of COVID-19 were found in Türkiye, according to statistics reported until September 2022, and around 6 per thousand of these cases resulted in mortality (JHU CSSE 2022). Istanbul is one of the most populous metropolitan areas in Europe, with around 15.9 million inhabitants (TSI 2021). The University of Health Sciences—Kanuni Sultan Süleyman Research and Training Hospital, situated within the borders of Küçükçekmece, one of Istanbul's districts with the highest population density, has been in operation since 1952 and continues to serve as a pandemic hospital as the number of patient cases across the nation rises. The experimental data of the study area includes the postal address information of 925 patients who applied to the hospital's COVID-19 services and polyclinic unit between March 11, 2020, and August 11, 2020 (Fig. 1). Additionally, the geographic locations of the official COVID-19 patients' data were provided from the Istanbul Metropolitan Municipality City Map for comparison after the geocoding process (URL-2). Pursuant to Turkish Personal Data Protection Law No. 6698, only the postal address information was used in this study, except for the identity information and other details of the patients.

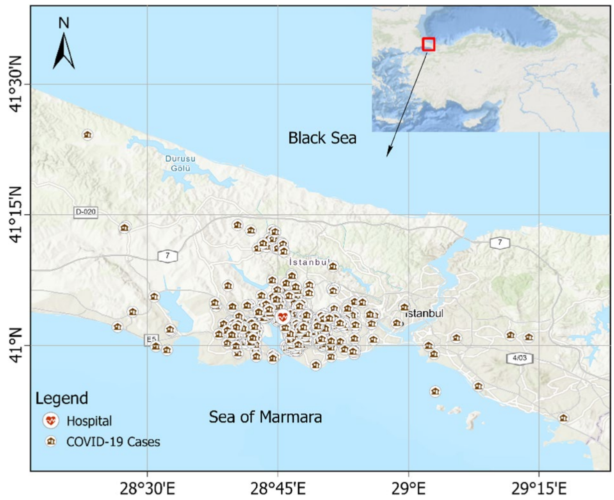

Fig. 1 Study area

3.2 Data collection

The University of Health Sciences—Kanuni Sultan Süleyman Research and Training Hospital COVID-19 services and Polyclinic unit, which is responsible for the development of official COVID-19 disease data, provided the data used in this study. The official COVID-19 data postal addresses were initially utilized as input for geocoding operations. The geocoding services of four worldwide map platforms, namely ArcGIS Online, Bing Maps, Google Maps, and HERE Maps, retrieved geographic coordinates and postal addresses in their standard forms during the preprocessing phase (Table 1).

After that, a comprehensive evaluation was conducted to assess official data and geocoding services' outputs, focusing on positional accuracy and address similarity. The authors established a binary relationship between reference data and retrieved addresses, comparing them to determine spatial alignment and semantic consistency (precision of address matching levels). The encoded associations were used as dependent variables to train and fine-tune machine learning algorithms. The evaluations and procedures are summarized in Table 2, providing an overview of the results obtained from the analysis of official and retrieved data.

3.3 Generation of input features

The effectiveness of different geocoding services has been analyzed in accordance with a set of input features that can be utilized with or without machine

Table 1 A sample of the online geocoding services results and benchmark data

Services	Postal address	Latitude and longitude
ArcGIS online	İnönü, İkizler Sokak 6, 34295, Küçükçekmece, İstanbul	[41.02746, 28.80212]
Bing maps	İkizler Sokak 6, 34295 Kucukcekmece	[41.02749, 28.80212]
Google maps	İnönü, İkizler Sk. No:6, 34295 Küçükçekmece/İstanbul	[41.02747, 28.80212]
HERE maps	İkizler Sokak 6, 34295, İnönü, Küçükçekmece, Istanbul	[41.02746, 28.80213]
Reference	İnönü Mah. İkizler Sk. No:6, 34295 Küçükçekmece/İstanbul	[41.02745, 28.80213]

Table 2 Precision of address matching levels and their detailed information for the machine learning process

Precision level	Description	Dependent variable encoding
Level 1	Exact address (neighborhood, street name, street number, district and province name)	1 (True)
Level 2	Otherwise	0 (False)

learning. Before the process of machine learning, general evaluation criteria such as positional accuracy and match rate between geocoding services and reference data were evaluated. On the other hand, seventeen various text similarity techniques were utilized to identify address similarities between each service and the reference data for the machine learning process. Measuring word similarity is one of the core processes utilized in a variety of tasks, such as title and toponym matching for geographical information retrieval (Santos et al. 2018), document plagiarism, information retrieval, automatic essay scoring, etc. (Gomaa and Fahmy 2013). This methodology uses lexical or semantic ways to differentiate between words. Lexical similarity is the relationship between two-character strings, whereas semantic similarity is the link between two characters with the same meaning but distinct syntactic features (Kilic and Gülgen 2020b). The detection of lexical similarity between two strings is carried out character-based, term-based, and hybrid-based with multiple similarity features. This study examined address similarities between each service and reference data using nine character-based, six term-based, and two hybrid text similarity methods (Table 3).

3.4 Machine learning classifiers for evaluation of address accuracy

To identify matching and non-matching address pairings in each service, the authors trained seven machine learning models with independent variables generated using seventeen distinct text similarity methods. Additionally, semantic similarity in the geocoding process is not taken into account in this study because it is not appropriate to measure it on its own. The percentages of similarity between the benchmark and retrieved addresses are independent variables in the machine learning procedure. Then, seven distinct machine learning approaches were implemented (Table 4).

Logistic Regression (LR) (Cox 1958) is a robust statistical model that generates probabilities along with classification results and is typically used for binary classification of linearly separable datasets, whereas Support Vector Machine (SVM) (Cortes and Vapnik 1995) seeks to determine the margins between classes based on the geometrical properties of a given dataset. Random Forest (RF) (Breiman 2001) is a collection of decision trees designed to reduce the variance of a learning model by (i) resampling random samples, (ii) constructing decision trees for each sample, and (iii) voting on predictions, also known as the bagging method. Gradient boosting machines (GBMs) (Friedman 2001) are used in machine learning to increase the generalization of a model by maximum correlating with the negative gradient of the loss function (Natekin and Knoll 2013; Bentéjac et al. 2021). In Adaptive Boosting (AdaBoost) (Freund and Schapire 1997), each weak learner and training data sample is weighted iteratively, and many weak learners are trained until the entire training data fits without any modification to the error function. eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin 2016) is a highly scalable ensemble learning classifier that overcomes the constraints of AdaBoost and GBM by regulating the overfitting-prone advanced regularizations L1 and L2. To speed up the processing time of XGBoost, a Microsoft team developed a generalized and efficient form of XGBoost called Light Gradient Boosting Machine (LightGBM), which lowered the sample

Table 3 The seventeen text similarity algorithms for machine learning process

Text similarity algorithms		
Character-based	Term-based	Hybrid-based
Levenshtein (Levenshtein 1966)	Dice- Sørensen (Dice 1945; Sørensen 1948)	Monge-Elkan (Monge and Elkan, 1997)
Jaro (Jaro 1989)	Jaccard (Jaccard 1901)	Soft-TFIDF (Doan et al 2012)
Jaro-Winkler (Winkler 1990)	Cosine (Huang 2008)	
Unigram	Bag Distance (Bartolini et al. 2002)	
Bigram	Overlap (Vijaymeena and Kavitha 2016)	
Trigram	Twersky (Twersky 1977)	
Needleman-Wunsch (Needleman and Wunsch 1970)		
Smith-Waterman (Smith and Waterman 1981)		
Gotoh (Gotoh 1982)		

size and feature size during training without significantly sacrificing accuracy and efficiency (Ke et al. 2017). Classifiers based on machine learning rely on factors to create predictions. Each variable influences the classification outcome, and the ML classifier selects the most significant factors. In this study, feature reduction strategies were not used to determine the most informative text similarity metric groups for the address matching quality of geocoding services. Model tuning was conducted using the grid-search technique, and the optimal configurations are presented in Table 4.

In the train/test procedure for ML classifiers, a tenfold stratified cross-validation method was employed. Receiver operator characteristics (ROC) curves were generated and mean area under curve (AUC) scores were calculated for each classifier and

Table 4 Utilized machine learning algorithms and most accurate hyperparameter configurations defined by grid-search

Classifier	Parameters
Logistic regression	
Support vector machines	Kernel = 'linear', regularization parameter = 1.0, degree of the polynomial kernel function = 3
Random forest	Number of estimators = 400, criteria = 'gini index',
Adaptive boosting	Number of estimators = 400, learning rate = 0.05
Gradient boosting machines	Number of estimators = 400, learning rate = 0.01, Max. depth = 8
Extreme gradient boosting machines	Number of estimators = 400, learning rate = 0.01, Max. depth = 8
Light gradient boosting machines	Max. depth = 8, learning rate = 0.05, number of leaves = 20, number of estimators = 400

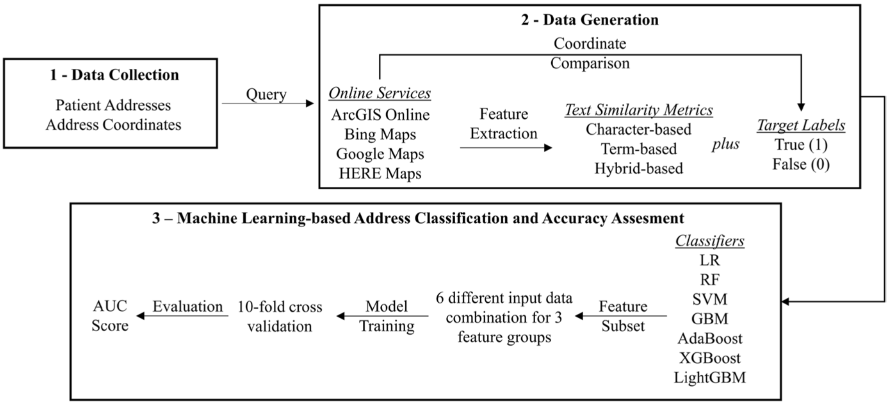

Fig. 2 Flowchart of the study

fold as well (Fig. 2). Hardware configuration was used for the training of ML models, which is Intel ® Core ™ i5-1035G4 CPU @ 1.10 GHz, 8 GB RAM.

4 Results and discussion

4.1 Overall evaluation results without machine learning

The general error distances of all geocoded addresses produced by each service are demonstrated in Fig. 3. For all points ($n=925$), the positional accuracy distributions between four different online geocoding services and the reference data were calculated.

As illustrated in Fig. 3, only around 80% of the geocoded points from the three other services, excluding Bing Maps, have error distances of less than 100 m. In contrast, about 33% of the geocoded points in Bing Maps have error distances of more than 1000 m. However, when the overall results in Fig. 3 are analyzed, it is revealed that all services have outlier distances. Table 5 presents the descriptive statistics of geocoded results for each service, including outliers.

Examining the descriptive statistics reported in Table 5 reveals that the positional accuracy of each service, as indicated by the medians of error distances, ranges from three to 14 m. Considering the matching rate, ArcGIS Online, Bing Maps, Google Maps, and HERE Maps services achieved 79.4%, 51.6%, 79.7%, and 82.9% success rates, respectively. Although it is seen that the HERE Maps service provides the best result compared to other services, Google Maps provided a shorter error distance based on the 95th percentile. In addition, Bing Maps has the lowest quality of match rate (51.6%), as well as the worst positional accuracy of all services.

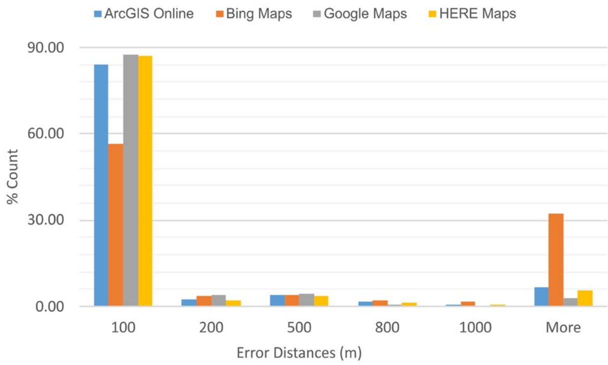

Fig. 3 The distribution of error distance calculated between each geocoded point and benchmark data for four services

Table 5 The descriptive statistics results of online geocoding services

Services	Matching rate (%)	Positional accuracy (m)							
		Min	Max	Median	Mean	STD	Percentile		
							75th	90th	95th
ArcGIS online	79.4	0.0	39,824.9	3.4	388.4	2,587.5	10.1	426.1	1,253.5
Bing maps	51.6	0.3	1,111,265.5	14.0	7,844.1	61,824.8	2,505.0	7,171.6	20,201.8
Google maps	79.7	0.1	21,377.6	6.2	172.8	1,117.1	15.5	153.3	329.1
HERE maps	82.9	0.0	37,929.2	2.8	284.2	1,853.1	8.0	245.7	1,100.6

4.2 Evaluation with machine learning and text similarity algorithms

In order to determine the optimal machine learning model for effective address matching, the performance of seven machine learning models was assessed using seventeen similarity criteria. To compare the effect of different types of text similarity metrics in geocoding services' accuracy assessment, a classification process was performed for each metric group as follows: (i) character, (ii) term, (iii) character + term, (iv) character + hybrid, (v) term + hybrid, and (vi) character + term + hybrid-based (Table 6). Among text similarity methods, there are two hybrid-based metrics known as Monge-Elkan and Soft-TFIDF. Since the number of hybrid-based metrics used is insufficient for the training/testing procedure of machine learning methods, they could not be taken into account as a group.

Figures 4, 5, 6, 7, 8, 9 illustrate the results with the highest accuracy (AUC) (Table 6) among SVC, LR, XGBoost, LightGBM, AdaBoost, GBM, and RF classifiers compared to geocoding services implementing text similarity methods. In character-based text similarity metrics, the RF classifier outperformed other classifiers with 0.8483, 0.8925, 0.8180, and 0.8249 AUC scores for ArcGIS Online, Bing Maps, Google Maps, and Here Maps, respectively (Fig. 4).

In term-based text similarity metrics, for ArcGIS Online and HERE Maps, the GBM classifier obtained the most accurate results with 0.8120 and 0.8112 AUC scores, whereas LightGBM and RF performed 0.7608 and 0.7256 AUC scores for Bing Maps and Google Maps, respectively (Fig. 5).

After the inclusion of both character and term-based text similarity metrics separately, by merging these types of text similarity metrics, the RF classifier achieved the highest AUC scores with 0.8299, 0.9017, 0.8338, 0.8228 for ArcGIS Online, Bing Maps, and Google Maps, and HERE Maps, respectively (Fig. 6).

Subsequently, involving both character and hybrid-based text similarity groups RF classifier performed 0.8533, 0.9061, 0.8256, and 0.8251 AUC scores by the RF classifier for ArcGIS Online, Bing Maps, Google Maps, and HERE Maps (Fig. 7).

Term and hybrid-based text similarity metrics were included together and the RF classifier performed 0.8198, 0.8228, 0.7920, and 0.8256 AUC scores and outperformed other classifiers likewise, for ArcGIS Online, Bing Maps, Google Maps, and HERE Maps, respectively (Fig. 8).

As a final process, with the inclusion of all features, the RF classifier outperformed other classifiers with 0.8577, 0.9084, 0.8240, and 0.8457 AUC scores for ArcGIS Online, Bing Maps, Google Maps, and HERE Maps, respectively (Fig. 9).

In terms of AUC values, the ROC curves depicted in Figs. 4, 5, 6, 7, 8, 9 reveal the optimal ML classifiers that are compatible with various grouped similarity measures. Despite lowering AUC scores with Fold 2 for ArcGIS Online and HERE Maps, and Fold 9 for Google Maps, mean AUC scores are still higher than match rate of services: 79.4, 51.6, 79.7, and 82.9% for ArcGIS Online, Bing Maps, and Google Maps, respectively (Table 2). Thus, this finding shows the benefits of machine learning classifiers over conventional methods for distinguishing between correct and erroneous addresses.

Results shown in bold typeface in Table 6 indicate the best ML classifiers that are compatible with various grouped similarity metrics used to determine address matching scores between four geocoding services and benchmark data. From this point of view, it can be seen that combining the capabilities of various grouped similarity metrics using ML is more successful than general evaluation findings containing only service matching scores for matching and non-matching addresses (Table 5 and Table 6). In addition, it is revealed that a similar pattern is observed in terms of ML classifiers for four online geocoding services, with the exception of the outcomes of term-based metrics. Among the seven kinds of ML algorithms, RF was the most successful classifier, ranging from 0.7919 to 0.9084 AUC scores for determining the matching and non-matching addresses (Table 6).

Further, among the six kinds of grouped text similarity metrics, it is also observed that only the process of term-based metric results (between 0.7256 and 0.8119 AUC scores) is more inconsistent than the others (Table 6). It is important to note that the RF classifier achieved lower AUC scores than GBM for ArcGIS Online (0.8119) and HERE Maps (0.8112) and LightGBM (0.7607) for Bing Maps services. Since term-based algorithms depend on completely matched token pairs, this inconsistency is caused by the lack of standardization between the address formats returned

Table 6 Mean area under curve (AUC) scores obtained with the different machine learning classifiers using various grouped similarity metrics (expressions in bold indicate the highest classification scores)

Text similarity metric group	Service	Classifier						
		SVC	LR	XGBoost	LightGBM	AdaBoost	GBM	RF
Character-based	ArcGIS online	0.8096	0.7944	0.8456	0.838	0.7874	0.7971	**0.8483**
	Bing maps	0.8294	0.7505	0.8846	0.8868	0.8557	0.8399	**0.8924**
	Google Maps	0.7127	0.6437	0.7911	0.7915	0.76	0.7821	**0.818**
	HERE maps	0.8065	0.7249	0.7993	0.8163	0.7606	0.8137	**0.8248**
Term-based	ArcGIS online	0.7517	0.8063	0.765	0.7928	0.769	**0.8119**	0.7762
	Bing maps	0.7214	0.5418	0.7369	**0.7607**	0.7531	0.7578	0.7559
	Google maps	0.6287	0.5683	0.7116	0.7009	0.6881	0.7094	**0.7256**
	HERE maps	0.7384	0.664	0.7482	0.7683	0.7515	**0.8112**	0.7831
Character and Term-based	ArcGIS online	0.8073	0.8044	0.8033	0.8109	0.785	0.8165	**0.8299**
	Bing maps	0.8362	0.7481	0.8846	0.8932	0.8612	0.8686	**0.9016**
	Google maps	0.7302	0.6458	0.8095	0.8134	0.7925	0.7715	**0.8337**
	HERE maps	0.8213	0.7272	0.8015	0.7987	0.7555	0.7987	**0.8228**
Character and hybrid-based	ArcGIS online	0.8273	0.7818	0.8203	0.8199	0.7788	0.7926	**0.8533**
	Bing maps	0.8408	0.754	0.8883	0.8871	0.8456	0.8523	**0.906**
	Google Maps	0.7434	0.6698	0.8221	0.8016	0.7432	0.7855	**0.8256**
	HERE Maps	0.807	0.725	0.8124	0.8147	0.7661	0.7909	**0.8251**
Term and hybrid-based	ArcGIS online	0.7958	0.7895	0.7863	0.8052	0.7642	0.808	**0.8197**
	Bing maps	0.7337	0.5785	0.8079	0.8034	0.7561	0.7517	**0.8228**
	Google maps	0.7191	0.6655	0.7695	0.7704	0.6988	0.6953	**0.7919**
	HERE maps	0.7936	0.6647	0.8002	0.8039	0.7456	0.8102	**0.8256**
Character, term and hybrid-based	ArcGIS online	0.8192	0.7948	0.8373	0.8274	0.7898	0.8306	**0.8576**
	Bing maps	0.8387	0.7496	0.8929	0.8969	0.8714	0.8421	**0.9084**
	Google maps	0.7474	0.6645	0.8031	0.7968	0.7316	0.7651	**0.8239**
	HERE maps	0.8192	0.7248	0.8162	0.8124	0.7717	0.8166	**0.8456**

Abbreviations: *SVC* Support vector classifier, *LR* Logistic regression, *XGBoost* Extreme gradient boosting machines, *LightGBM* Light gradient boosting machines, *AdaBoost* Adaptive boosting, *GBM* Gradient boosting machines, *RF* Random forest

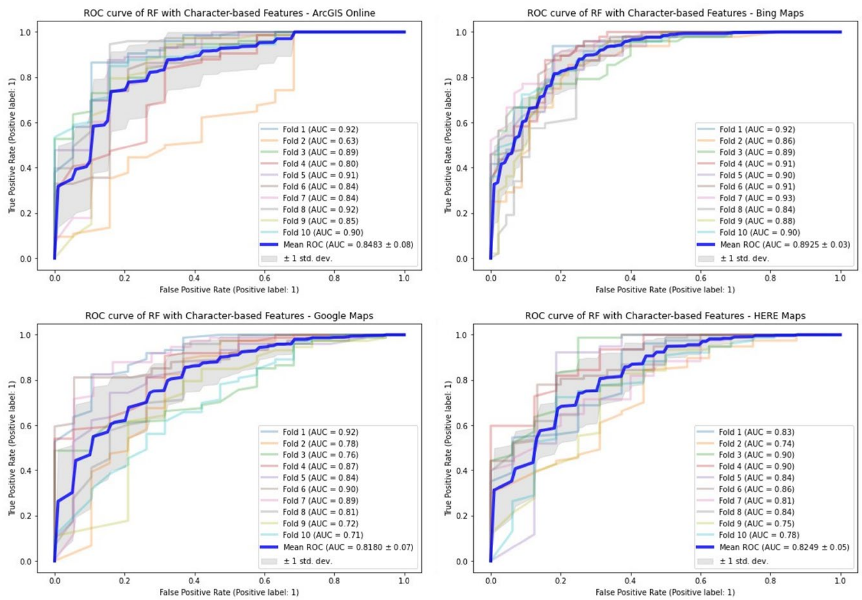

Fig. 4 Area under curve (AUC) scores of random forest (RF) classifier, Character (ROC receiver operator characteristics; Upper left: ArcGIS Online, Upper right: Bing Maps, Lower left: Google Maps, Lower right: HERE Maps)

by the services and the reference addresses. A similar situation emerged as a result of hybrid-based algorithms. Looking at the findings of the hybrid-based metrics, it is seen that the AUC scores of the character + hybrid-based algorithms (ranging from 0.8251 to 0.9060) are higher than those obtained with the term + hybrid-based algorithms (ranging from 0.7919 to 0.8256) for all services in Table 6.

As one can see in Table 6, the highest AUC scores (between 0.8239 and 0.9084) for all services, with the exception of Google Maps (0.8337 AUC score), with character- and term-based similarity metrics, are attained with a combination of three separate similarity measures. On the other hand, despite the fact that the results obtained using term-based algorithms are inferior to all other possibilities, it is evident that these algorithms have a positive effect on the combination of all metric data.

4.3 Limitations and future research directions

This study was carried out utilizing a dataset comprising 925 postal addresses provided by individuals in isolation as a result of COVID-19. At this point, considering the number of both postal address data and the real-world geography it represents, it is necessary to acknowledge that our analysis is limited to a certain specific region. Nevertheless, the outcomes of this study undeniably bear significance for other analogous research domains and address datasets, particularly within the context of

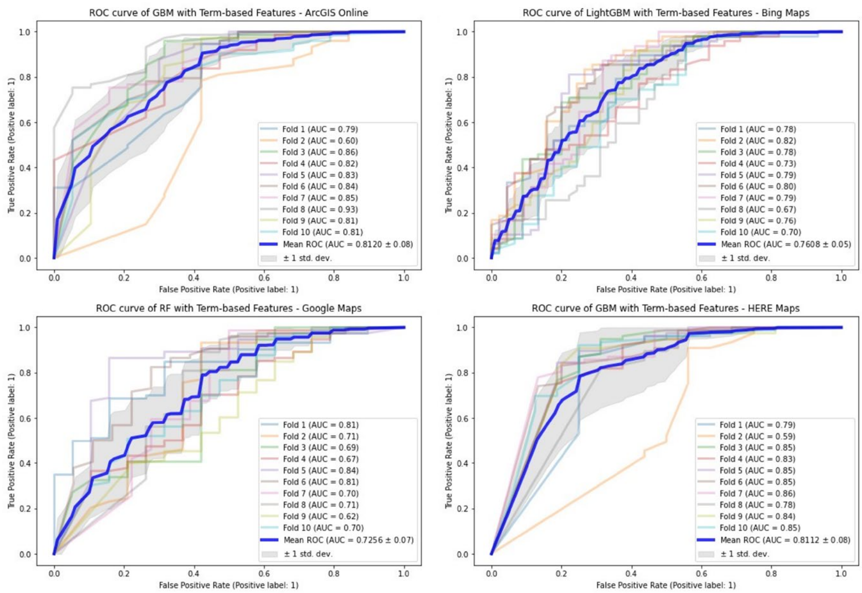

Fig. 5 Area under curve (AUC) scores of gradient boosting machines (GBM), LightGBM and random forest (RF) classifier, Term (ROC receiver operator characteristics; Upper left: ArcGIS Online, Upper right: Bing Maps, Lower left: Google Maps, Lower right: HERE Maps)

Türkiye. Although address formats retrieved from different online geocoding services and the resulting outputs may vary, the insights derived from our research indicated that the RF method can detect reliable address matching in widely-used online geocoding services in Türkiye.

The performance of the address matching process in this study is limited to the Turkish addressing system. It is important to note that due to the presence of different formats and rules in addresses across different countries and even within regions of a single country, the findings obtained cannot be generalized to address formats of other countries (Koumarelas et al. 2018; Lee et al. 2020; Kilic and Gülgen 2020a). These differences complicate the geocoding process for global map platforms and have an adversely impact on the accuracy of the processes. In order to evaluate the global effectiveness of online geocoding services, it may be necessary for future research to utilize analyses based on datasets with varying formats. This would help in understanding and improving the performance of these services.

Some online geocoding services lack some of the standard components of the Turkish addressing system. For example, the Bing Maps service does not provide its users with data for the address component "neighborhood". This has a negative effect on the accuracy of address matching results produced by machine learning techniques. Standardizing multiple components, such as street types or avenues, neighborhoods, door numbers, postal codes, districts, provinces, and established abbreviations, becomes essential in order to enhance and supplement the results of machine learning (e.g., by incorporating additional rules).

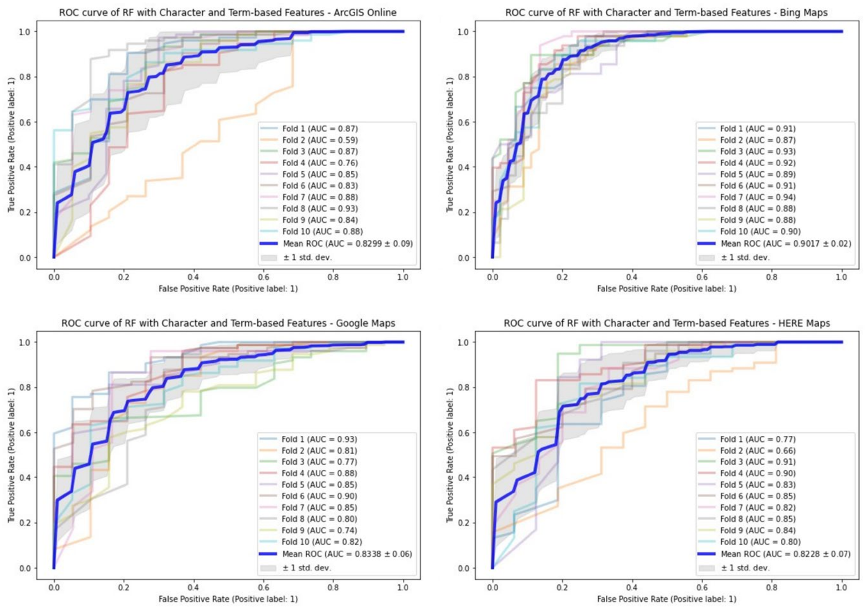

Fig. 6 Area under curve (AUC) scores of random forest (RF) classifier, Character + Term (ROC receiver operator characteristics; Upper left: ArcGIS Online, Upper right: Bing Maps, Lower left: Google Maps, Lower right: HERE Maps)

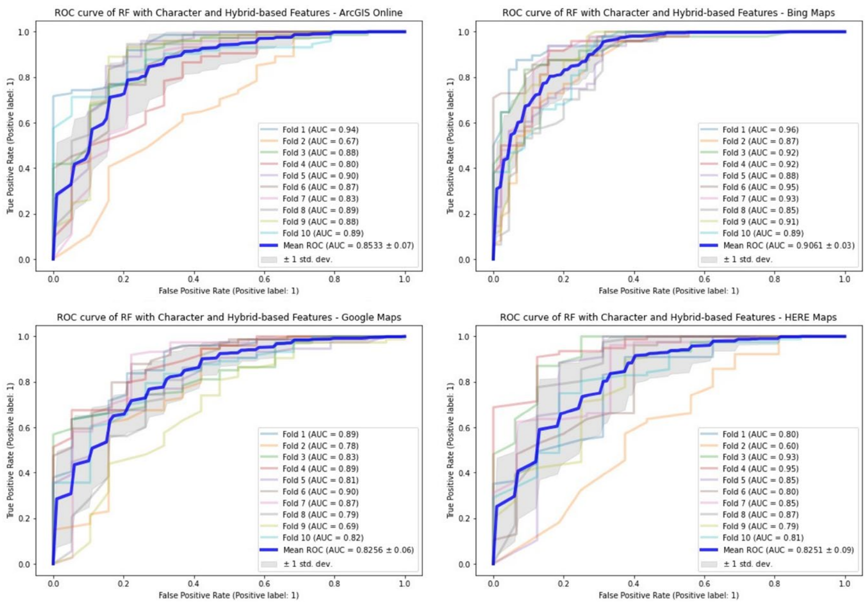

Fig. 7 Area under curve (AUC) scores for random forest (RF) classifier, Character + Hybrid (ROC receiver operator characteristics; Upper left: ArcGIS Online, Upper right: Bing Maps, Lower left: Google Maps, Lower right: HERE Maps)

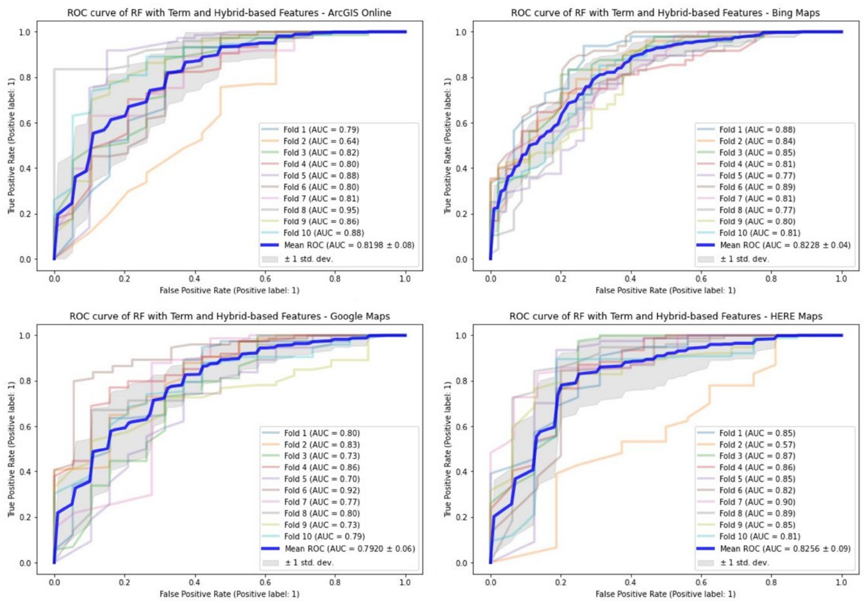

Fig. 8 Area under curve (AUC) scores for random forest (RF) classifier, Term + Hybrid (ROC receiver operator characteristics; Upper left: ArcGIS Online, Upper right: Bing Maps, Lower left: Google Maps, Lower right: HERE Maps)

Lastly, in compliance with the Turkish Personal Data Protection Law No. 6698, only the address and location information of COVID-19 patients were utilized. It was impossible to conduct various analyses based on parameters such as age, gender, height, weight, or the presence of chronic diseases among the reported patients. However, these parameters were not necessary for the scope of this study. Depending on whether similar laws exist in other countries, the impacts of the laws on the results obtained can be investigated and the generalizability of the approach proposed in this study to systems in other countries can be assessed.

5 Conclusion

With the qualification that geocoding is an important topic in health science, it is necessary to reveal that precise determination of residential addresses of persons with reportable communicable diseases to contain outbreaks and to identify and manage contacts. In this context, the empirical comparison of four different online geocoding services offered by ArcGIS Online, Bing Maps, Google Maps, and HERE Maps is presented for matching and non-matching addresses using both general evaluation methods and machine learning classifiers.

The evaluation results without machine learning reveal several differences among these four services. In general, Bing Maps produced lower matching rates and larger

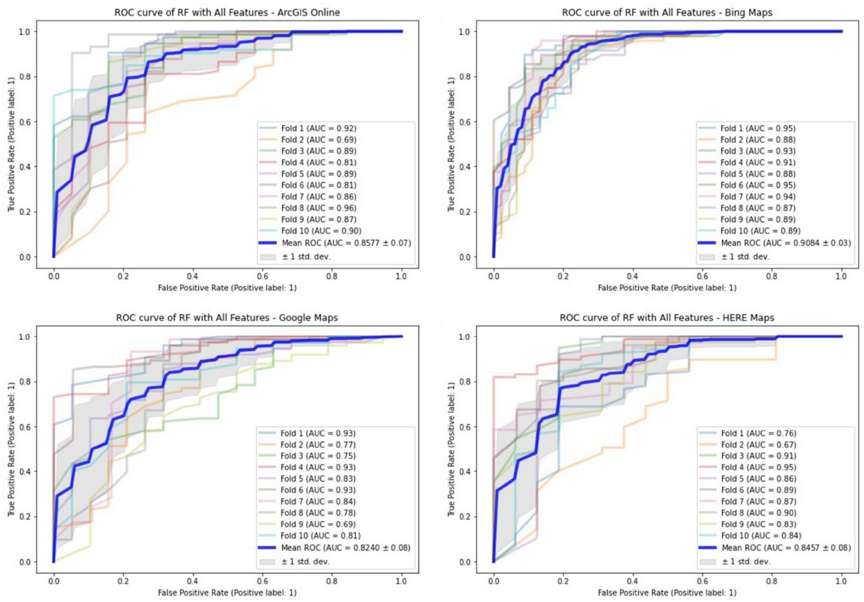

Fig. 9 Area under curve (AUC) scores for random forest (RF) classifier, all features (ROC receiver operator characteristics; Upper left: ArcGIS Online, Upper right: Bing Maps, Lower left: Google Maps, Lower right: HERE Maps)

error distances than the others. On the other hand, although HERE Maps provided more accurate points among these services, it is not achieved the same success in terms of positional accuracy. In addition, all services except Google Maps produced less accurate results for the 95th percentile.

Furthermore, the feasibility of detecting address accuracy is examined using machine learning classifiers, at actively used geocoding services in daily life. Selected machine learning classifiers were employed on the postal addresses of COVID-19 patients, and as input features, different text similarity metrics were generated based on address pairs from the geocoding service and reference addresses. Our experiments revealed that machine learning classifiers, especially RF classifier, is capable of assessing the accuracy of geocoding services and usage in circumstances such as global crises, and those classifiers can smoothly be adapted without powerful hardware specifications for daily usage. In the meantime, to reduce the processing time of the employed classifiers and identify the most informative similarity metric group, we separately included them in the machine learning classification process. Although it has been observed that term-based metrics decrease accuracy in separate and double groups, the most successful results were achieved by including entirely different types of similarity metrics. It is clear from the proceedings that each classifier and similarity metric have their advantages and disadvantages; however, the RF classifier outperformed other classifiers with the inclusion of all similarity metrics.

Our study can guide the researchers who aim to investigate and increase the accuracy of the geocoding services as well. Moreover, it can be applied to various datasets from different countries where the task of matching address elements presents challenges. In accordance with their precision of address matching levels, researchers, practitioners, and stakeholders will be able to utilize online geocoding services more effectively and dependably in this way. Future studies will focus on increasing the accuracy of the geocoding service through extended address data (multiple reference data) and state-of-the-art methods based on deep neural network-based architectures.

Acknowledgements The authors would like to thank the University of Health Sciences Kanuni Sultan Süleyman Research and Training Hospital's professionals who indirectly contributed to this study. The Scientific Research Platform of the Ministry of Health and the Institutional Review Board of the University of Health Sciences – Kanuni Sultan Süleyman Research and Training Hospital approved this retrospective study (No: 2020.06.80) on May 23, 2020 and June 18, 2020, respectively.

Author's contribution BK: Conceptualization, Writing—Original Draft, Methodology. OCB: Methodology, Writing—Original Draft, Visualization. FG: Writing—Reviewing and Editing, Supervision. MG: Data collection, Funding acquisition. PA: Data collection, Funding acquisition.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval and consent to participate Not applicable.

Consent for publication Not applicable.

References

AkakbaA, Lahmar B (2020) The use of geocoding for home healthcare application and management an epidemic situation. Case of COVID-19 virus outbreak. Geogr Pannon 24(4)

Andresen MA, Malleson N, Steenbeek W, Townsley M, Vandeviver C (2020) Minimum geocoding match rates: an international study of the impact of data and areal unit sizes. Int J Geogr Inf Sci 34(7):1306–1322

ARS (2007) Adres Kayıt Sistemi Uygulama Yönergesi [Address registration system application directive]. https://adres.nvi.gov.tr/Home

Bartolini I, Ciaccia P, Patella M (2002) String matching with metric trees using an approximate distance. In: International symposium on string processing and information retrieval (pp 271–283). Springer.
Bentéjac C, Csörgő A, Martínez-Muñoz G (2021) A comparative analysis of gradient boosting algorithms. Artif Intell Rev 54(3):1937–1967
Bichler G, Balchak S (2007) Address matching bias: ignorance is not bliss. Policing Int J Police Strateg Manage 30(1):32–60
Breiman L (2001) Random forests. Mach Learn 45(1):5–32
Briz-Redón Á, Martinez-Ruiz F, Montes F (2020) Reestimating a minimum acceptable geocoding hit rate for conducting a spatial analysis. Int J Geogr Inf Sci 34(7):1283–1305
Chen T, Guestrin C (2016). Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp 785–794)
Cohen SE, Stookey J, Anderson N, Morris D, Singzon T, Dann M, Chen CC (2022) Using geocoding to identify COVID-19 outbreaks in congregate residential settings: San Francisco's outbreak response in single-room occupancy hotels. Public Health Reports, 00333549221128301
Comber S, Arribas-Bel D (2019) Machine learning innovations in address matching: a practical comparison of word2vec and CRFs. Trans GIS 23(2):334–348
Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
Cox DR (1958) The regression analysis of binary sequences. J Roy Stat Soc: Ser B (methodol) 20(2):215–232
Cui Y (2013) A systematic approach to evaluate and validate the spatial accuracy of farmers market locations using multi-geocoding services. Appl Geogr 41:87–95
Davis CA, Fonseca FT (2007) Assessing the certainty of locations produced by an address geocoding system. GeoInformatica 11(1):103–129
Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26(3):297–302
Doan A, Halevy A, Ives Z (2012) Principles of data integration. Morgan Kaufmann
Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat, pp 1189–1232
Goldberg DW, Cockburn MG (2012) The effect of administrative boundaries and geocoding error on cancer rates in California. Spatial Spatio Temp Epidemiol 3(1):39–54
Goldberg DW, Jacquez GM, Mullan N (2013) Geocoding and health. Geographic health data: fundamental techniques for analysis, London, CAB International, pp 51–71Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. International Journal of Computer Applications, 68(13), 13–18.
Gomaa WH, Fahmy AA (2013) A survey of text similarity approaches. Int J Comput Appl 68(13):13–18
Gotoh O (1982) An improved algorithm for matching biological sequences. J Mol Biol 162(3):705–708
Huang A (2008) Similarity measures for text document clustering. In: Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008) (vol 4, pp 49–56)
Jaccard P (1901) Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin De La Société Vaudoise Des Sciences Naturelles 37(142):547–579
Jaro MA (1989) Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J Am Stat Assoc 84(406):414–420
JHU CSSE (2022) Johns Hopkins University Center for Systems Science and Engineering. COVID-19 Dashboard. https://coronavirus.jhu.edu/map.html. Accessed on 01 Sept 2022
Karabegovic A, Ponjavic M, Hukic M (2021) Epidemic location intelligence system as response to the COVID-19 outbreak in Bosnia and Herzegovina. Appl Syst Innov 4(4):79
Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Liu TY (2017) Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30
Khalifa MB, Diaz Redondo RP, Vilas AF, Rodríguez SS (2017) Identifying urban crowds using geo-located Social media data: a Twitter experiment in New York City. J Intell Inf Syst 48(2):287–308
Kilic B, Gülgen F (2020a) Accuracy and similarity aspects in online geocoding services: a comparative evaluation for Google and Bing maps. Int J Eng Geosci 5(2):109–119
Kilic B, Gülgen F (2020b) Investigating the quality of reverse geocoding services using text similarity techniques and logistic regression analysis. Cartogr Geogr Inf Sci 47(4):336–349
Kilic B, Gülgen F, Kocataş A, Gürtürk M, Abay P, Yılmaz Y, Anataca G (2022) COVID-19 hastalarının izlenmesinde çevrimiçi coğrafi kodlama servislerinin karşılaştırmalı analizi. Jeodezi Ve Jeoinformasyon Dergisi 9(2):95–107
Koumarelas I, Kroschk A, Mosley C, Naumann F (2018) Experience: enhancing address matching with geocoding and similarity measure selection. J Data Inf Qual (JDIQ) 10(2):1–16

Kounadi O, Lampoltshammer TJ, Leitner M, Heistracher T (2013) Accuracy and privacy aspects in free online reverse geocoding services. Cartogr Geogr Inf Sci 40(2):140–153
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
Lee K, Claridades ARC, Lee J (2020) Improving a street-based geocoding algorithm using machine learning techniques. Appl Sci 10(16):5628
Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions, and reversals. Soviet Phys Doklady 10(8):707–710
Levine N, Kim KE (1998) The location of motor vehicle crashes in Honolulu: a methodology for geocoding intersections. Comput Environ Urban Syst 22(6):557–576
Lin Y, Kang M, Wu Y, Du Q, Liu T (2020) A deep learning architecture for semantic address matching. Int J Geogr Inf Sci 34(3):559–576
Matci DK, Avdan U (2018) Address standardization using the natural language process for improving geocoding results. Comput Environ Urban Syst 70:1–8
Monge AE, Elkan CP (1997) An efficient domain-independent algorithm for detecting approximately duplicate database records. In: Proceedings of the SIGMOD 1997 workshop on research issues on data mining and knowledge discovery
Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot 7:21
Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453
PTT (2013) Adres Yazım Standartları [Address writing standards]. http://postakodu.ptt.gov.tr/Dosyalar/adres.pdf
Qin X, Parker S, Liu Y, Graettinger AJ, Forde S (2013) Intelligent geocoding system to locate traffic crashes. Accid Anal Prev 50:1034–1041
Rashidian S, Dong X, Jain SK, Wang F (2018) EaserGeocoder: integrative geocoding with machine learning (demo paper). In: Proceedings of the 26th acm sigspatial international conference on advances in geographic information systems (pp 572–575)
Ratcliffe JH (2004) Geocoding crime and a first estimate of a minimum acceptable hit rate. Int J Geogr Inf Sci 18(1):61–72
Roongpiboonsopit D, Karimi HA (2010) Comparative evaluation and analysis of online geocoding services. Int J Geogr Inf Sci 24(7):1081–1100
Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL (2006) Geocoding in cancer research: a review. Am J Prev Med 30(2):S16–S24
Sah, S. (2020). Machine learning: a review of learning types. 2020070230. https://doi.org/10.20944/preprints202007.0230.v1
Santos R, Murrieta-Flores P, Martins B (2018) Learning to combine multiple string similarity metrics for effective toponym matching. Int J Digital Earth 11(9):913–938
Simon A, Singh M (2015) An overview of M learning and its Ap. Int J Electrical Sci Electrical Sci Eng IJESE 22
Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195–197
Sørensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biologiske Skirfte 5:1–34
TSI (2021) Turkish Statistical Institute website. https://www.tuik.gov.tr/. Accessed on 18 Aug 2022
Tversky A (1977) Features of similarity. Psychol Rev 84(4):327–352
URL-1 (2022) T.C. Sağlık Bakanlığı COVID-19 Bilgilendirme Platformu [Republic of Turkey Ministry of Health COVID-19 Information Platform], https://covid19.saglik.gov.tr/. Accessed on 24 Aug 2022
URL-2 (2022) İstanbul Şehir Haritası [Istanbul Metropolitan Municipality City Map]. https://sehirharitasi.ibb.gov.tr/. Accessed on 16 July 2022
Vijaymeena MK, Kavitha K (2016) A survey on similarity measures in text mining. Mach Learn Appl Int J 3(2):19–28
Winkler WE (1990) String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In: Proceedings of the section on survey research methods of the American statistical association, ASA
Yildirim V, Yomralioglu T, Nisanci R, Inan H (2014) Turkish street addressing system and geocoding challenges. Proc Inst Civil Eng Municipal Eng 167(2):99–107
Yin Z, Ma A, Goldberg DW (2019) A deep learning approach for rooftop geocoding. Trans GIS 23(3):495–514

Zandbergen PA (2008) A comparison of address point, parcel and street geocoding techniques. Comput Environ Urban Syst 32(3):214–232

Zandbergen PA (2009) Geocoding quality and implications for spatial analysis. Geogr Compass 3(2):647–680

Authors and Affiliations

Batuhan Kilic[1] · Onur Can Bayrak[1] · Fatih Gülgen[1] · Mert Gurturk[1] · Perihan Abay[2]

✉ Batuhan Kilic
batuhank@yildiz.edu.tr

Onur Can Bayrak
onurcb@yildiz.edu.tr

Fatih Gülgen
fgulgen@yildiz.edu.tr

Mert Gurturk
mgurturk@yildiz.edu.tr

Perihan Abay
perihan.abay@saglik.gov.tr

1 Department of Geomatic Engineering, Yildiz Technical University, Esenler, Istanbul 34220, Türkiye

2 University of Health Sciences, Kanuni Sultan Süleyman Research and Training Hospital, Küçükçekmece, Istanbul 34303, Türkiye