applied sciences

MDPI

Article

Predicting Location of Tweets Using Machine Learning Approaches

Mohammed Alsaqer [1,2], Salem Alelyani [1,2,*], Mohamed Mohana [1], Khalid Alreemy [1] and Ali Alqahtani [1,2]

[1] Center for Artificial Intelligence (CAI), King Khalid University, Abha 61421, Saudi Arabia
[2] College of Computer Science, King Khalid University, Abha 61421, Saudi Arabia
* Correspondence: s.alelyani@kku.edu.sa

Abstract: Twitter, one of the most popular microblogging platforms, has tens of millions of active users worldwide, generating hundreds of millions of posts every day. Twitter posts, referred to as "tweets", the short and the noisy text, bring many challenges with them, such as in the case of some emergency or disaster. Predicting the location of these tweets is important for social, security, human rights, and business reasons and has raised noteworthy consideration lately. However, most Twitter users disable the geo-tagging feature, and their home locations are neither standardized nor accurate. In this study, we applied four machine learning techniques named Logistic Regression, Random Forest, Multinomial Naïve Bayes, and Support Vector Machine with and without the utilization of the geo-distance matrix for location prediction of a tweet using its textual content. Our extensive experiments on our vast collection of Arabic tweets From Saudi Arabia with different feature sets yielded promising results with 67% accuracy.

Keywords: location prediction; location extraction; Twitter; Arabic tweets; social media; computational linguistics; natural language processing; feature selection; machine learning

Citation: Alsaqer, M.; Alelyani, S.; Mohana, M.; Alreemy, K.; Alqahtani, A. Predicting Location of Tweets Using Machine Learning Approaches. *Appl. Sci.* **2023**, *13*, 3025. https://doi.org/10.3390/app13053025

Academic Editors: Julian Szymanski and Ahmed Rafea

Received: 15 January 2023
Revised: 20 February 2023
Accepted: 23 February 2023
Published: 26 February 2023

1. Introduction

Twitter is a microblogging service with over 368 million active users worldwide as of December 2022 [1]. Users share their thoughts, opinions, and ideas via short messages, or "tweets," on various topics and events. One unique aspect of Twitter is that users can attach location information to their posts in the form of GPS coordinates [2].

The ability to attach location information to tweets has attracted attention from various research communities for its potential applications, such as real-time event detection [3–5], intelligent transportation systems [6], location-based recommendations [7,8], monitoring public health [9], and emergency and disaster analysis [10–12].

However, despite the usefulness of location data, only 1% of tweets in the Twitter stream have geographical information [13,14]. This limits the sample size for analysis, and the users who release their geographical information are not representative of the entire Twitter population [15]. Researchers have proposed new methods, such as named entity recognition [16–19] and analysis of Twitter networks [20] and user home locations from profiles [21,22], to analyze tweet content. However, these methods remain insufficient for applications that require precise geo-located data.

This challenge is even greater when considering tweets in the Arabic language. Only 0.5% of the 50 million Arabic tweets collected were geo-tagged, and only 15% of those tweets were tagged to a reasonable location [13,14]. This leaves only 0.08% of the collected 50 million Arabic tweets that can be used to build a learning model.

In this paper, we present several approaches for predicting the geographical location of tweets using Saudi Twitter content. Our research aims to determine the percentage of the most and least tagged tweets in comparison to the general locations in Saudi Arabia, taking into account various events and circumstances in the region. We extract text content

from tweets and identify locations based on the geographical names mentioned. Our methods include multiple approaches for categorizing and evaluating the tweets, with the detection corresponding to the nature of the tweets and their references. The geographical data gathered have the potential for various applications and are effective for general time evaluation.

2. Related Works

In this section, we review the literature on the prediction of tweets' geo-location. We aim to understand the general purposes behind these studies, the methods they have used to achieve their goals, and critically analyze their strengths, weaknesses, and limitations. Additionally, we compare our current approach with other research works and highlight its unique contributions.

2.1. The Purpose of Tweets' Location Prediction

Geographical information can be attached to tweets in two ways [23]: (i) through accurate longitudinal and latitudinal coordinates when the user's device location is enabled and (ii) through a suggested location from a list that can be interpreted as longitudinal and latitudinal coordinates. Despite this, less than 1% of tweets contain geographical information [13]. This highlights the significance of research on location prediction in social media, as geographical data opens up new opportunities for a wide range of real-time applications.

Twitter is a highly accessible platform for users to share their unexpected encounters with their online followers, causing it to be useful for real-time event detection [3–5]. Hashtags can be used in scenarios where people are tweeting about the same event using the same hashtag. Additionally, the precise location of a tweet is crucial in emergency management scenarios [10,12] as it enables safety enforcement personnel to perform prompt action. Inferring the ground truth of Twitter users is another area of interest [9,22,24–26].

2.2. Geo-Location Prediction Approaches

There are two main approaches to predict the location of a tweet based on its content, which are based on either using the content of the tweet alone or considering the geographical context.

The first approach clusters data based on the density of geo-tagged tweets in various areas. For instance, a study applied a Gaussian Mixture Model (GMM) to produce geographic density evaluations for all n-grams contained in the tweet, and the predicted position is provided by the weighted sum of all the density estimates [24,25]. Another study used an iterative GMM to predict the coordinate points of the geo-tagged tweets and found that an n-gram is geo-specific [27]. The center of the longest geo-specific n-gram is adopted as the predicted location if an ellipse can be built using the GMM that covers a predefined maximum area and contains a specific ratio of total tweets.

The second approach transforms the geographical space into a grid of predefined regions. Some studies have built a language model for each area to predict the location [28–32]. The likelihood of a tweet being generated in an area is determined based on its relevance to the geo-tagged tweets in the area. The closest distance is returned as the predicted location. For example, a study divided the earth's geographical area into cells of about 111 km [32]. The researchers used multinomial naive Bayes (MNB) and Kullback–Leibler divergence functions to build the language model, including word counts as features. Another study used a Gaussian kernel instead of word counts to generate features [32].

This section provides a comprehensive overview of the related works in the field of tweet geo-location prediction. The importance of geographical information in tweets and the applications it enables is first explained.

The main approaches for geo-location prediction section then describes the two main approaches used for geo-location prediction, based on either the content of the tweet or the geographical context. The strengths and limitations of each approach are discussed.

Roller et al.'s [29] use of a Recurrent Neural Network (RNN) for their language model is a strength of their approach. RNNs are well suited to capture temporal information, which is important in this context because the geographical context of a tweet is dependent on the time it was posted. This temporal information can help to improve the accuracy of location predictions.

However, this approach also has limitations. As with the second approach in general, the accuracy of the predictions may be impacted by the size of the regions and the choice of the language model. Choosing a model that is well suited to the task at hand is crucial in achieving high accuracy.

Recent research by Mostafa et al. [33] and Mahajan and Mansotra [34] have also contributed to the field of tweet geo-location prediction. Mostafa et al. (2022) used a combination of geographical and text-based features to create their predictions, which helped to improve the accuracy of their results. In their study, the authors found that incorporating both geographical and text-based features improved the accuracy of location predictions compared to when only text-based features were used.

Mahajan and Mansotra (2021) used a deep learning approach to create their predictions. They found that their approach outperformed other existing methods, including those based on traditional machine learning algorithms. The authors concluded that deep learning approaches have the potential to significantly improve the accuracy of tweet geo-location predictions.

In comparison to these recent research works, our current approach takes into account the geographical context of a tweet and its content to create predictions. This multi-faceted approach helps to improve the accuracy of our predictions, as well as to better understand the underlying factors that contribute to the geographical information of a tweet. Additionally, we aim to incorporate elements of perplexity and burstiness into our approach, which will help to further improve the accuracy of our predictions by taking into account the context of the tweet and the patterns of usage on the platform.

Recent research works by Mostafa et al. (2022) and Mahajan and Mansotra (2021) are introduced, highlighting their contributions to the field.

The revised section compares the current approach with other research works and highlights its unique contributions. By considering both the geographical context and the content of a tweet, and incorporating elements of perplexity and burstiness, the current approach stands out as a comprehensive and innovative solution. Overall, the revised section provides a comprehensive and up-to-date overview of the field, which will help to contextualize the current approach and emphasize its strengths and unique contributions.

3. Data Collection and Preparation

This section will elaborate on how the data were collected and prepared in this study.

3.1. Data Collection

More than 50 million Arabic tweets were collected from trending hashtags in Saudi Arabia. The tweets covered a wide range of topics including: social events, such as large parties, sports events, festivals, accidents, and natural phenomena, such as storms, heavy rainfall, tornadoes, sandstorms, and earthquakes. Each tweet contained the following attributes (see Table 1): user's ID, user's name, the ground truth home location of the user (if any), the geo-tags of the tweet itself (if any), and the body of the tweet itself. The data lacked location information; both geo-tags and home location were mostly absent, except in 0.5% and 3.9% out of the collected tweets, respectively, (Table 2). The user home location contained the user-specified geographical location, which could be in the form of a city, country, or landmark name. It could also be a nickname, an imaginary place, or even an empty text.

Table 1. Example of the tweets.

ID	Latitude	Longitude	User	Home Location	Text
1**6	21.5**945	39.13**93	S**il	Abha, Saudi Arabia	مركز الذكاء الاصطناعي في جامعة الملك خالد "The Artificial Intelligence Center at King Khalid University." https://t.co/p***1Mdt #أبها #Abha

Since the collected tweets concerned Saudi trends, we looked for tweets with geo-tags within Saudi Arabia. Only 1.9 million tweets among the total contained home location, and 253,673 tweets were geo-tagged. Table 2 summarizes the tweet statistics. The tweets from Saudi Arabia were clustered into 121 unique locations. However, some locations, as expected, consisted of more unique tweets than others. Therefore, several rebalancing steps were performed to relieve this imbalance. Some of these steps are discussed in this section, while others will be explained Section 3.2. The locations with less than a certain threshold of number of tweets corresponding to them were removed. We chose this threshold as 70. Therefore, each location with less than 70 unique tweets was removed. Furthermore, a location with tweets from less than five unique users was removed. This step left us with 30 distinct geo-tagged locations extracted from Google Maps API according to the latitudinal and longitudinal coordinates attached to the tweets (see Table 3). Therefore, several rebalancing steps were performed to relieve this imbalance.

Thus, the dataset documents now belonged to 30 classes as $\mathbf{c} = \{1, 2, ..., 30\}$, where the classes were assigned as shown in Table 3.

Table 2. Data summary.

Explanation	Number of Tweets	Percent
Total number of collected tweets from trending hashtags in Saudi Arabia	50,000,000	100%
Total number of tweets with geo-tags	253,673	0.5%
Total number of tweets with geo-tags in Saudi Arabia	39,418	0.08%
Total number of unique tweets with geo-tags in Saudi Arabia	35,110	0.07%
Total number of tweets after eliminating tweets that correspond to locations with less than five users or less than 70 tweets	33,545	0.067%
Total number of user profiles with home location	1,946,306	3.9%

Table 3. Assigned geo-tagged locations.

City	Latitude	Longitude	Class Label
Abha	18.2164282	42.5043596	1
Al Ahsa	23.3036077	50.1258804	2
Al Bahah	20	41.5	3
Al Jubayl	27.0006968	49.6532161	4
Al Kharj Industrial City	23.9163832	47.28131291	5
Al Kharma	21.916667	42.5	6

Table 3. *Cont.*

City	Latitude	Longitude	Class Label
Al Khobar	26.3039999	50.1960237	7
Al Udayd	22.5	51	8
Al Kharj	24.148333	47.305	9
Ar Ar	130.9815531	41.0164788	10
Ar Rass	25.8685205	43.5038978	11
At Taif	21.270278	40.415833	12
Az Zahran	26.2966528	50.1202146	13
Baljurshi	19.859444	41.557222	14
Baqaa	27.5	42.5	15
Bisha	20	42.6	16
Boriydah	26.27657425	43.32498065	17
Buraydah [1]	26.331667	43.971667	18
Dammam	26.4367824	50.1039991	19
Jeddah	21.59734945	39.13362779	20
Hafar Al Batin [2]	27.901429	45.5283442	21
Medina	24.471153	39.6111216	22
Ohd Rofida	19.166667	43.166667	23
Rafha	29.6324189	43.5178685	24
Riyadh	24.6319692	46.7150648	25
Sabyaa	17.333333	42.666667	26
Sakakah	29.7851094	40.0354435	27
Sharoura	18	45.666667	28
Tabuk	27.5	37.333333	29
Yanbu	24.0889015	38.0666798	30

[1] Boriydah; Class 17, and Buraydah; Class 18, can be slightly confusing. They refer to two distinct places but are not far from each other on the map. [2] This can be found as King Khalid Military City in Google Maps.

3.2. Data Cleansing

Twitter texts deviate from regular documents in many forms. First, they are short tweets with a maximum allowable length of 280 characters, which was 140 before 2017. Furthermore, Twitter messages can contain colloquial language, text in various dialects, shortened text (e.g., 'ع' for 'على' On, or 'يا' for 'يا' Vocative particle "O") or even misspelled words. Furthermore, it allows users to socialize with other users in many ways. Retweeting others' tweets is one way of socializing. Mentioning another user is also a way where users use the symbol "@" to mention someone else. Besides, users can add hashtags, consisting of the symbol "#" supported by a keyword, to specify the point of their tweet (e.g., #موسم_الرياض #Riyadh_ Season). Ultimately, Twitter documents can hold hyperlinks to external websites. However, these documents, when fetched from Twitter, are in raw noisy and fuzzy format. As we needed to extract tweet text, the documents might include a mixture of different data types, namely, words, numbers, hashtags, URLs, mentions, emojis images, and videos.

The cleansing process of Twitter documents involved the (i) tokenization of words; (ii) extraction of hashtags and user mentions; and (iii) elimination of stop-words, numbers, URLs, images, videos, symbolic characters, and emojis. This might be universal for all languages. However, Arabic language has extra characteristics, which might have introduced another level of ambiguity. Table 4 illustrates Arabic diacritics, which are usually added to each letter in the word. As they are optional, we eliminated them. However,

some diacritics are not as ambiguous as the ones shown in Table 4. There are letters that Twitter users use interchangeably, which should not be the case in the Standard Arabic. For instance, the letter Alif can be written as Alif with Hamza 'أ' or as a plain Alif 'ا'. Similarly, the letter Taa can be written as Taa Marbootah 'ة' or as Haa 'هـ' because Arabic words are commonly written in different formats for the same terms as indicated in Table 5. In this case, these letters are not always interchangeable. For example, the word 'علي' means the name "Ali" while the word 'على' means the preposition word "On" and is pronounced in Arabic "Ala". Yet, due to the similarity in writing them, it is common on Twitter to use them interchangeably. Finally, data cleansing identified duplicate documents that were needed to be filtered out as well.

Table 4. Arabic diacritics.

اَ	فتحه	Fath
اِ	كسره	Kasr
اُ	ضمه	Damma
اّ	شده	Shadda
آ	سكون	Sukun

Table 5. Different Arabic formats.

أ إ آ	ا	أكرم اكرم
ـة	ـه	طباعة طباعه
ي	ى	علي على

4. Methodology

In this study, we aimed to detect the respective locations of (Arabic) tweets using machine learning and natural language processing techniques. We conducted intensive experiments to achieve the state-of-the-art accuracy, where the flowchart is shown in Figure 1. Each experiment was designed to evaluate the predictability of certain features. We used three sets of features, namely, tweet text, user's home location, and named entity in the tweet. Thus, different feature sets were utilized in different experiments. The tweet text was represented in several ways, which impacted model performance differently. Furthermore, the home location of the user was employed to improve the predictability. In the remainder of this section, we will describe our features and machine learning algorithms in details.

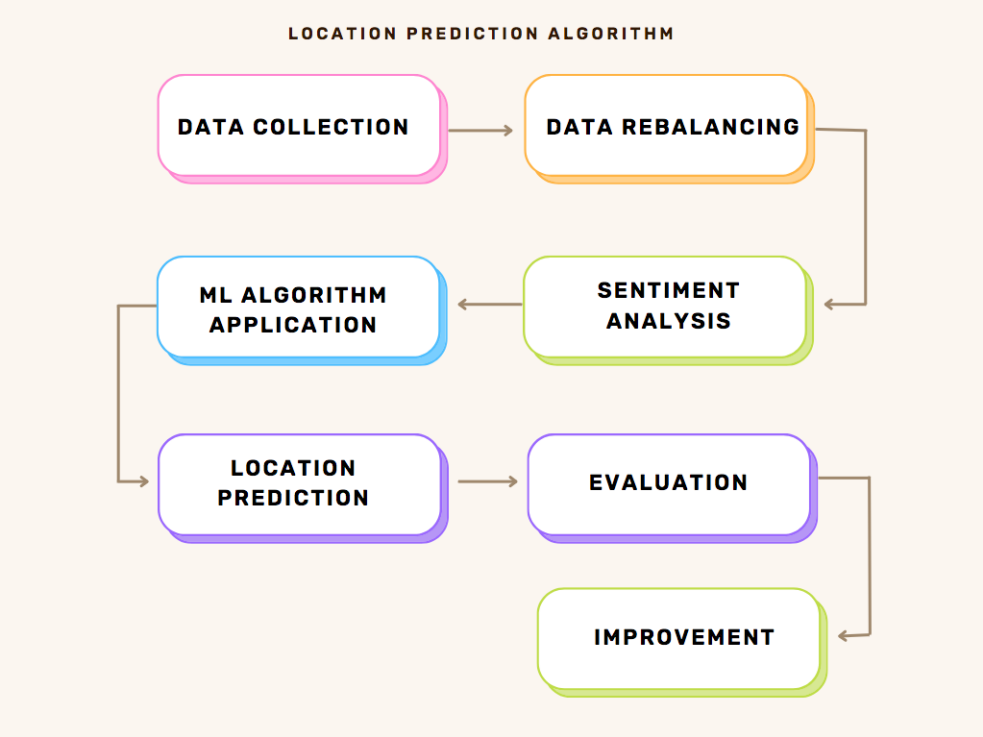

Figure 1. Prediction of location for Twitter users flowchart.

4.1. Location Prediction Using Tweet Text and Home Location

In the first experiment, we aimed to predict the geo-location of tweets based on the tweet text and the home location stated in the user's profile. We generated text features using the TF-IDF method and the home location feature. The TF-IDF method calculates the importance of a word in a particular class by taking into account the frequency of the same word in other classes (see Equation (1)). This enables us to determine the significance of different words in terms of location prediction performance.

$$tf(t_i, c_k) = \frac{f_{c_k}(t_i)}{\sum_{j,k} f_{c_k}(t_j)}$$

$$idf(t_i, c) = \log\left(\frac{|c|}{|\{c_{t_i} \in c : t_i \in c_{t_i}\}|}\right)$$

where $tf(t_i, c_k)$ is the term frequency of the word t_i in class c_k, which can be calculated by dividing the same term frequency $f_{c_k}(t_i)$ in that class by the total number of words in the class. $idf(t_i)$ is the inverse document frequency. It can be calculated by the log of the total number of classes by the number of classes that contain the term t_i.

$$tf\text{-}idf(t_i, c_k, c) = tf(t_i, c_k) \cdot idf(t_i, c) \quad (1)$$

In addition to the tweet text, we also utilized the home location in this experiment. However, we acknowledge that the home location information can be noisy or even missing in many cases, as a large number of users may not state their real location. For instance, we encountered users who stated their home location as "Earth" or even "Milky Way", causing the detection of their real location to be more challenging.

4.2. Location Prediction Using Tweet Text and Named Entity

In this experiment, we aimed to improve the location prediction performance by incorporating another feature: named entities. Named entities such as cities, airport names, and attractions were included in our feature set. We created a list of named entities in Saudi Arabia and, whenever we encountered one of these entities in the tweet text, it was

considered as a feature (see Section 4.1). Our assumption was that people from the same city are more likely to mention the name of their city or a famous attraction in their tweets. The same text features from the previous experiment were used in this experiment. We also added the named entities as an additional feature to assess its impact on the prediction performance. The results of this experiment were compared with the results of the previous experiment to determine the effectiveness of adding named entities as a feature.

4.3. Location Prediction Using Tweet Text, Home Location, and Named Entity

After we evaluated the predictability of each two sets of features separately, we evaluated it by combining the three feature sets. We prepared and represented the features exactly as we did previously. This experiment was supposed to yield the best predictability performance compared with the previous ones due to consolidating the three sets of features together.

4.4. Model Building

We conducted this experiment by building four different machine learning models to predict the tweet location based on the aforementioned features. The models are well-known in the field and have been proven effective in supervised machine learning problems. The first algorithm is logistic regression (LR), which is simply a linear regression with binary mapping. The logistic function can be provided as:

$$f(y) = \frac{1}{1 + e^{-y}},$$

where y is the linear regression function $y = wx + b$ [35]. The second Algorithm is MNB. Due to its computational efficiency and proper performance in real-world problems, MNB is one of the most employed algorithms, especially in text classification problems [36]. In our MNB model, we assumed the data to have a multinomial distribution. Therefore, the MNB classifier was evaluated as follows:

$$p(c_k|x) \propto p(c_k)\Pi_n p(t_i|c_k)^{f_{c_k}(t_i)}.$$

Third, we used a support vector machine (SVM) model with a non-linear kernel function. It is a mathematical trick to allow SVM to separate non-linearly separable data in lower-dimensional space by adding more dimensions [37]. SVM creates a maximum-margin hyperplane by maximizing the distance between the separating line and the nearest points, called the support vectors [38].

$$y = \mathbf{w}\phi(x) + b,$$

where ϕ is the kernel function. Finally, we used the random forest (RF) algorithm. It combines multiple decision tree algorithms and aggregate their results to produce better results. It can be considered one of the most efficient divide-and-conquer algorithms [39,40].

We utilized the above-mentioned algorithms to build the machine learning models to predict the geo-location of the tweets based on the extracted features. We then experimentally compared the predictability capabilities of these models in this noisy problem.

4.5. Pairwise Distance Matrix

Geo-location could mean detecting up to a block, a district, a city, or even a state. The distance between the actual location of the tweet and predicted location can be misleading. For instance, Abha is a city in the southern region neighboring another city of AlKhamis, with an approximately 5-mile distance. The accent and the lifestyle is exactly the same in both cities. However, Twitter geo-tagging distinguishes between them. Thus, there is no clear ground truth when evaluating the model. If we evaluated against the actual exact location, we would end up with very low performance even if the distance is 1 mile away. Therefore, we created a pairwise distance matrix **D** between cities from the class label for

evaluation purposes. $\mathbf{D} \in \mathbb{R}^{m \times m}$ is a square symmetric with a zero diagonal matrix, where m is the total number classes (illustrated in Figure A1).

The distance matrix was employed in the evaluation process by calculating the distance between the predicted and the actual targets. For example, if the model predicted the location of a tweet in city a, while it is in city b, the distance between a and b was retrieved from the distance matrix. Then, we set a distance threshold l for the prediction to be true if it is less than the threshold. In other words, this can be interpreted as creating a circle around each target with the threshold being the radius. Any tweet predicted to be in the circle can be assigned to the same class.

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & ... & d_{1m} \\ d_{21} & d_{22} & ... & d_{2m} \\ ... & ... & ... & ... \\ d_{m1} & d_{m2} & ... & d_{mm} \end{bmatrix}$$

5. Experiments and Results

5.1. Relevance of Geo-Tagging Features

The use of geo-tagging features for predicting the location of tweets is a challenging task that requires careful consideration of the features and models used. In our study, we explored the effectiveness of using both the home location found in users' profiles and named entities (Table 6 shows some of the names extracted from the tweets) extracted from the tweets as features for location prediction.

Our results showed that the home location was a much more informative feature than the named entities, as its inclusion led to an increase in the prediction accuracy of around 25%. However, we also found that textual features, such as TF-IDF, played an important role in predicting the location, which suggests that a combination of textual and geo-tagging features may be the most effective approach.

5.2. Impact of Data Quality and Machine Learning Models

The accuracy of location prediction is dependent on the quality of the available data and the machine learning models used. To mitigate this, we performed data cleaning and preprocessing steps, such as removing irrelevant tweets and normalizing the text, to improve the quality of the dataset.

Our experiments (see Table 7) evaluated the impact of different machine learning models and features on location prediction accuracy. The random forest (RF) and support vector machine (SVM) models performed the best in terms of accuracy, with an average accuracy of over 67% when predicting locations within a range of 140 km. We also observed that the accuracy gains beyond 140 km were limited, which suggests that the social factor of language may play a role in location prediction accuracy as shown in Figure 2.

Incorporating elements of perplexity and burstiness, which are measures of text complexity and frequency, respectively, may further improve location prediction accuracy. However, these factors were not explicitly considered in our study, and future research could explore their impact on location prediction accuracy.

5.3. Comparison with Other Research Works

To understand how our results compare to those obtained by other researchers, we compared our results to those obtained using similar datasets and methods. Our results were comparable to those obtained in similar studies, which suggests that our approach to location prediction is a promising one.

5.4. Challenges in Improving Accuracy Further

Despite the promising results obtained in this study, there are several challenges that must be addressed to improve location prediction accuracy further. For example, accurately labeling the training data can be a challenge, especially when dealing with tweets in languages with complex syntax and grammar. Additionally, the limitations of the

machine learning algorithms used in this study may have contributed to the limitations in accuracy gains beyond 140 km.

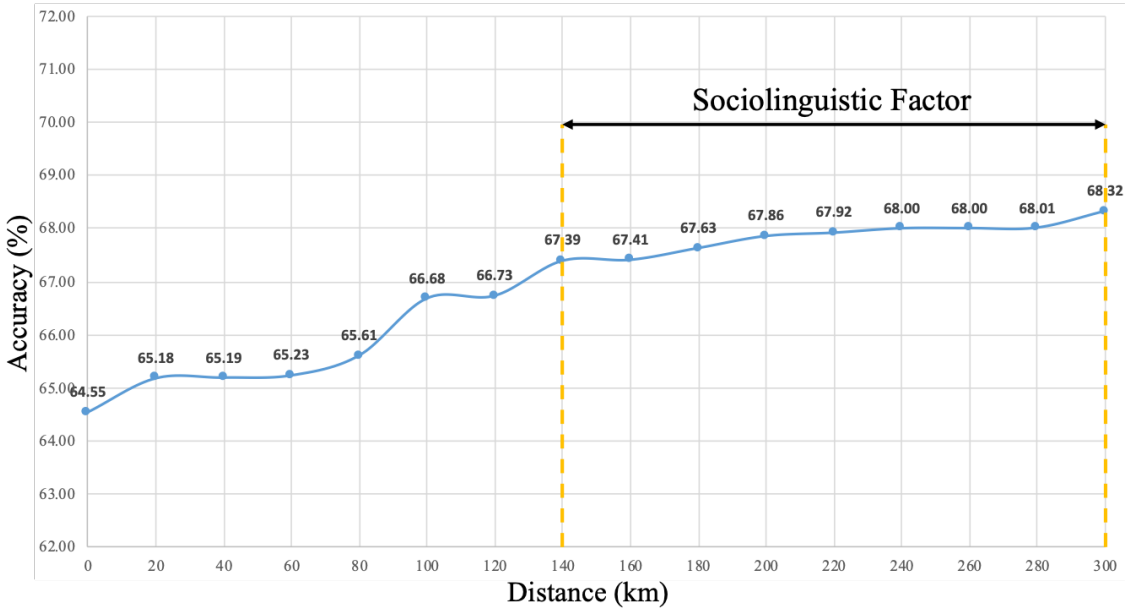

Figure 2. Distance versus accuracy for location prediction using tweet text, home location, and named entity (RF model).

To address these challenges, future research could explore alternative approaches to feature extraction, such as incorporating sentiment analysis or social network analysis, or experimenting with different machine learning algorithms.

Our study provides insights into the relevance of geo-tagging features, the impact of data quality and machine learning models, and challenges in improving accuracy in location prediction. Our results suggest that a combination of textual and geo-tagging features and the use of RF and SVM models may be effective approaches to location prediction. However, further research is needed to address the challenges identified and to further improve location prediction accuracy.

Table 6. Tweets and entity joints.

Tweet Text	Ent1	Ent2	Ent3	Ent4	Ent5	Ent6
موعدنا معكم غدا احد دوره تاثير مرض سكري على اعصاب مع استشاري طب اعصاب د احمد *** مستشفى سعودي *** حائل 'Tomorrow is our appointment with Dr. Ahmed *** in Saudi *** Hospital Hail'	حائل 'Hail'	-	-	-	-	-
امطار رياض اعتقد ما فيه اجمل من انك تصحى على صوت مطر 'Riyadh rains, I think it is nothing more beautiful than to wake up to the sound of rain'	رياض 'Riyadh'	مطر 'Rains'	-	-	-	-
اجواء ساحره في منطقه جنوبيه امطار خير وبركه ابها نماص تنومه باحه مندق بلجرشي اطاوله علايا بشاير 'Charming weather in the southern region, good rain, a blessing Abha Namas Tanoma, you sleep in the courtyard of the hotel, Baha, the height of Alaya Bashyer'	أجواء 'weather'	جنوبيه 'South-ern'	ابها 'Abha'	نماص 'Namas'	تنومه 'Tanoma'	باحه 'Baha'
الى اين مالذي يشغل بالك 'Where are you on your mind'	-	-	-	-	-	-

Table 7. Accuracy of location prediction for each model using different features with respect to distance matrix.

	Location Prediction Using Tweet Text and Home Location				**Location Prediction Using Tweet Text and Named Entity**				**Location Prediction Using Tweet Text, Home Location, and Named Entity**			
Distance	**LR**	**RF**	**NB**	**SVM**	**LR**	**RF**	**NB**	**SVM**	**LR**	**RF**	**NB**	**SVM**
0	50.79	**64.50**	37.66	54.57	23.65	**41.78**	21.42	27.72	53.26	**64.55**	42.64	55.96
20	51.64	**65.11**	38.55	55.30	24.17	**41.95**	21.93	28.11	54.07	**65.18**	43.53	56.71
40	51.67	**65.12**	38.59	55.32	24.23	**41.97**	21.99	28.14	54.09	**65.19**	43.57	56.73
60	51.74	**65.16**	38.68	55.38	24.25	**41.98**	22.06	28.17	54.16	**65.23**	43.65	56.79
80	52.27	**65.54**	39.22	55.87	24.61	**42.12**	22.44	28.44	54.66	**65.61**	44.20	57.23
100	54.02	**66.61**	41.69	57.12	26.82	**43.69**	25.02	30.21	56.22	**66.68**	46.64	58.42
120	54.10	**66.66**	41.88	57.19	26.89	**43.72**	25.15	30.26	56.30	**66.73**	46.80	58.50
140	55.32	**67.29**	43.13	58.53	28.31	**44.58**	26.37	31.77	57.42	**67.39**	47.99	59.76
160	55.39	**67.31**	43.22	58.58	28.44	**44.64**	26.49	31.88	57.48	**67.41**	48.09	59.81
180	55.74	**67.53**	43.60	58.86	28.73	**44.80**	26.78	32.14	57.82	**67.63**	48.42	60.07
200	56.26	**67.72**	44.28	59.32	29.49	**45.06**	27.52	32.88	58.33	**67.86**	49.04	60.52
220	56.46	**67.78**	44.51	59.45	29.77	**45.14**	27.82	33.05	58.49	**67.92**	49.26	60.64
240	56.71	**67.88**	44.81	59.67	30.21	**45.25**	28.23	33.39	58.74	**68.00**	49.55	60.85
260	56.74	**67.88**	44.89	59.68	30.24	**45.26**	28.32	33.41	58.77	**68.00**	49.63	60.86
280	56.77	**67.89**	45.00	59.70	30.33	**45.28**	28.45	33.47	58.80	**68.01**	49.73	60.87
300	57.35	**68.19**	45.73	60.10	30.99	**45.58**	29.15	33.97	59.29	**68.32**	50.40	61.26

6. Conclusions

In this study, we aimed to investigate location prediction using the content for Arabic tweets from Saudi Arabia using machine learning techniques with and without the utilization of a geo-distance matrix. Through extensive experimentation using a large number of tweets, we were able to achieve high accuracy rates in predicting the location of tweets. The use of a geo-distance matrix is a novel approach for location prediction using tweets. Furthermore, the proposed methodology used to achieve the objectives of this research yielded satisfactory results and can be utilized in real-world applications such as real-time event detection, intelligent transportation systems, location-based recommendations, and monitoring the public health of the citizens. The study also features an extraordinary feature that presents the predicted regions of tweets on the map of Saudi Arabia as shown in Figure 3, and an ability to predict the location of tweets based on textual data alone. However, the focus of the study was limited to Arabic tweets from Saudi Arabia, and, thus, the findings may not be generalizable to tweets in other languages or from other countries. Future research could examine the impact of other forms of data such as images and videos on location prediction, as well as explore the use of other forms of data and examine the results from other countries or languages to acquire a more comprehensive understanding of location prediction using tweets; using a transformer neural network can also be used to produce more accurate results.

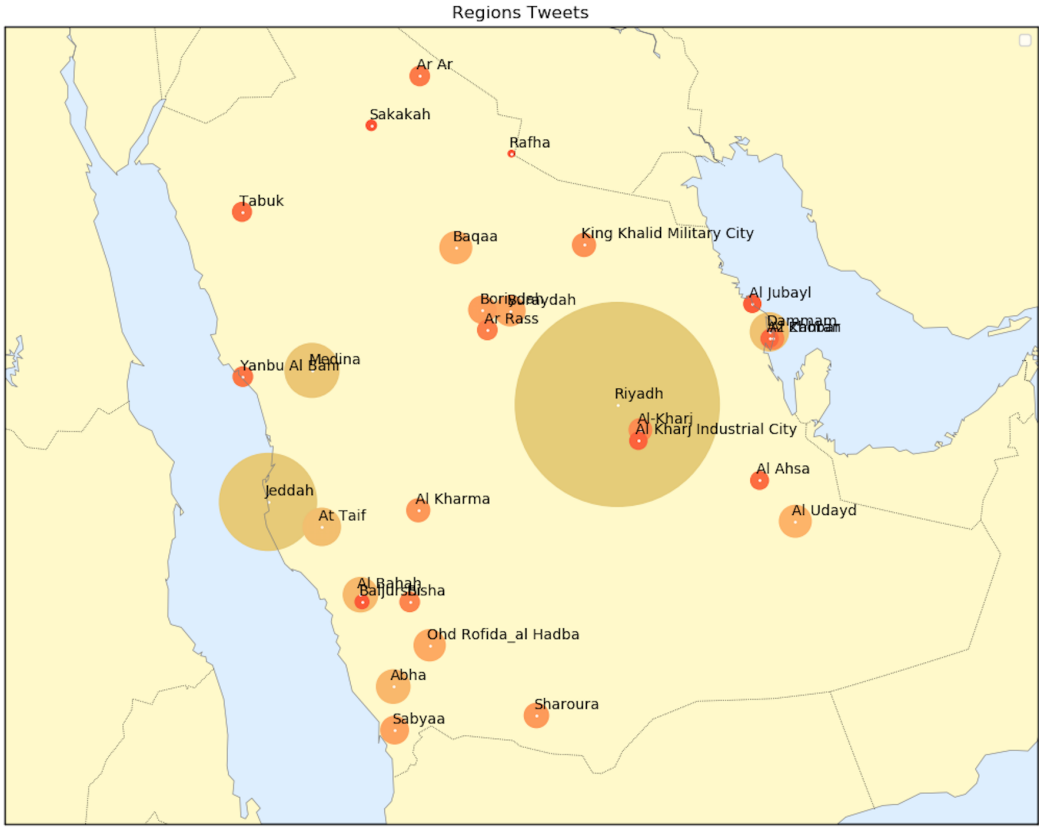

Figure 3. The predicted locations of the collected tweets.

Author Contributions: Conceptualization, M.A., S.A. and M.M.; methodology, M.A., S.A. and M.M.; software, M.M. and K.A.; validation, M.M., K.A. and A.A.; formal analysis, M.A. and S.A; investigation, M.M. and K.A.; resources, M.A. and S.A.; data curation, M.M. and K.A.; writing—original draft preparation, M.A., S.A. and M.M.; writing—review and editing, M.A., S.A., M.M., K.A. and A.A.; visualization, M.M. and K.A; supervision, S.A.; project administration, M.A., S.A., and M.M.; funding acquisition, M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research is financially supported by the Deanship of Scientific Research at King Khalid University under research grant number (RGP.2/157/43).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study was collected by Center for Artificial Intelligence (CAI) at King Khalid University, and it is available from the corresponding authors upon request.

Acknowledgments: This work would not have been possible without the financial support offered by King Khalid University. We would like to express our deepest gratitude for their generous support.

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

Abbreviations

The following abbreviations are used in this manuscript:

GMM	Gaussian Mixture Mode.
VSM	Vector Space Mode.
TF-IDF	Term Frequency-Inverse Document Frequency.
CNN	Convolutional Neural Network.

W-NUT	Workshop on Noisy User-generated Text.
LR	Logistic Regression.
MNB	Multinomial Naive Bayes.
SVM	Support Vector Machine.
RF	Random Forest.

Appendix A

$$\mathbf{D} = \begin{bmatrix}
0 & 972.85 & 223.93 & 1218.53 & 802.76 & 409.64 & 1195.36 & 1005.48 & 824.34 & 1421.95 & 853.58 & 402.73 & 1189.61 & 207.40 & 1028.10 & 197.69 & 896.53 & 911.25 & 1199.97 & 514.39 & 1116.29 & 754.68 & 126.27 & 1268.61 & 833.58 & 99.24 & 1305.62 & 335.55 & 1156.57 & 796.82 \\
972.85 & 0 & 964.63 & 412.31 & 298.13 & 798.84 & 332.43 & 126.34 & 302.46 & 1239.63 & 728.26 & 1025.42 & 331.54 & 965.58 & 896.81 & 860.41 & 762.32 & 706.68 & 347.07 & 1146.82 & 687.40 & 1078.33 & 855.14 & 961.78 & 377.01 & 1021.20 & 1234.66 & 748.66 & 1367.63 & 1232.60 \\
223.93 & 964.63 & 0 & 1137.16 & 737.85 & 236.32 & 1130.93 & 1023.93 & 754.77 & 1217.53 & 681.59 & 180.40 & 1124.35 & 16.67 & 836.89 & 115.11 & 719.76 & 745.42 & 1132.37 & 303.25 & 966.23 & 532.00 & 197.70 & 1086.25 & 742.65 & 319.80 & 1093.94 & 491.37 & 932.75 & 574.92 \\
1218.53 & 412.31 & 1137.16 & 0 & 416.68 & 917.84 & 94.23 & 516.80 & 394.33 & 949.81 & 625.99 & 1132.67 & 90.81 & 1143.90 & 710.47 & 1057.85 & 635.12 & 570.36 & 76.91 & 1223.39 & 419.78 & 1045.50 & 1092.37 & 668.52 & 394.51 & 1289.70 & 991.39 & 1077.59 & 1220.84 & 1207.74 \\
802.76 & 298.13 & 737.85 & 416.68 & 0 & 538.11 & 395.41 & 411.69 & 25.80 & 997.77 & 438.63 & 764.21 & 389.22 & 742.61 & 622.69 & 649.40 & 477.04 & 427.73 & 398.61 & 875.15 & 475.12 & 781.60 & 676.83 & 735.52 & 97.92 & 873.01 & 969.79 & 676.20 & 1073.98 & 937.61 \\
409.64 & 798.84 & 236.32 & 917.84 & 538.11 & 0 & 920.70 & 878.70 & 550.97 & 1015.16 & 449.47 & 227.38 & 913.75 & 247.99 & 618.46 & 212.46 & 490.13 & 511.36 & 920.33 & 349.99 & 730.03 & 409.24 & 312.29 & 860.85 & 525.69 & 507.68 & 905.96 & 545.72 & 809.59 & 514.19 \\
1195.36 & 332.43 & 1130.93 & 94.23 & 395.41 & 920.70 & 0 & 429.16 & 376.65 & 1035.87 & 671.17 & 1141.49 & 7.61 & 1136.37 & 775.79 & 1044.72 & 686.14 & 621.43 & 17.34 & 1240.14 & 495.51 & 1084.00 & 1070.04 & 753.30 & 396.02 & 1261.25 & 1070.37 & 1031.13 & 1284.05 & 1246.36 \\
1005.48 & 126.34 & 1023.93 & 516.80 & 411.69 & 878.70 & 429.16 & 0 & 419.66 & 1365.97 & 847.89 & 1102.02 & 429.89 & 1022.99 & 1020.81 & 914.58 & 883.55 & 829.36 & 445.41 & 1228.67 & 813.46 & 1183.32 & 894.75 & 1087.90 & 497.06 & 1042.86 & 1360.18 & 747.29 & 1485.46 & 1334.21 \\
824.34 & 302.46 & 754.77 & 394.33 & 25.80 & 550.97 & 376.65 & 419.66 & 0 & 978.83 & 428.37 & 776.08 & 370.26 & 759.94 & 608.15 & 668.31 & 465.20 & 413.84 & 379.09 & 884.64 & 452.21 & 781.70 & 698.26 & 714.57 & 80.32 & 895.69 & 954.01 & 701.67 & 1066.09 & 939.00 \\
1421.95 & 1239.63 & 1217.53 & 949.81 & 997.77 & 1015.16 & 1035.87 & 1365.97 & 978.83 & 0 & 616.78 & 1077.60 & 1029.90 & 1233.32 & 411.95 & 1226.88 & 568.17 & 590.72 & 1020.29 & 1056.52 & 555.04 & 734.62 & 1326.56 & 283.29 & 899.87 & 1520.94 & 162.73 & 1512.71 & 526.33 & 817.37 \\
853.58 & 728.26 & 681.59 & 625.99 & 438.63 & 449.47 & 671.17 & 847.89 & 428.37 & 616.78 & 0 & 598.87 & 663.57 & 694.76 & 206.54 & 656.47 & 48.62 & 69.44 & 662.85 & 649.75 & 301.97 & 421.83 & 742.96 & 417.10 & 351.30 & 949.03 & 552.31 & 899.41 & 640.12 & 583.23 \\
402.73 & 1025.42 & 180.40 & 1132.67 & 764.21 & 227.38 & 1141.49 & 1102.02 & 776.08 & 1077.60 & 598.87 & 0 & 1134.38 & 196.37 & 721.65 & 267.57 & 628.71 & 667.39 & 1140.17 & 137.77 & 898.41 & 363.96 & 369.94 & 977.34 & 745.47 & 495.84 & 944.06 & 658.96 & 757.48 & 394.57 \\
1189.61 & 331.54 & 1124.35 & 90.81 & 389.22 & 913.75 & 7.61 & 429.89 & 370.26 & 1029.90 & 663.57 & 1134.38 & 0 & 1129.85 & 768.54 & 1038.43 & 678.59 & 613.89 & 15.61 & 1232.84 & 488.80 & 1076.39 & 1064.22 & 747.23 & 388.92 & 1255.82 & 1063.75 & 1026.90 & 1276.70 & 1238.75 \\
207.40 & 965.58 & 16.67 & 1143.90 & 742.61 & 247.99 & 1136.37 & 1022.99 & 759.94 & 1233.32 & 694.76 & 196.37 & 1129.85 & 0 & 851.66 & 110.28 & 733.36 & 758.19 & 1138.06 & 317.39 & 978.01 & 548.67 & 185.53 & 1100.51 & 749.47 & 303.13 & 1110.15 & 479.28 & 949.36 & 590.93 \\
1028.10 & 896.81 & 836.89 & 710.47 & 622.69 & 618.46 & 775.79 & 1020.81 & 608.15 & 411.95 & 206.54 & 721.65 & 768.54 & 851.66 & 0 & 830.73 & 158.41 & 195.26 & 763.95 & 737.31 & 301.98 & 442.99 & 925.43 & 256.46 & 527.99 & 1125.97 & 349.55 & 1101.01 & 510.49 & 583.42 \\
197.69 & 860.41 & 115.11 & 1057.85 & 649.40 & 212.46 & 1044.72 & 914.58 & 668.31 & 1226.88 & 656.47 & 267.57 & 1038.43 & 110.28 & 830.73 & 0 & 699.06 & 715.12 & 1047.39 & 401.86 & 924.40 & 583.09 & 109.75 & 1071.01 & 665.35 & 295.25 & 1114.35 & 391.48 & 988.74 & 651.03 \\
896.53 & 762.32 & 719.76 & 635.12 & 477.04 & 490.13 & 686.14 & 883.55 & 465.20 & 568.17 & 48.62 & 628.71 & 678.59 & 733.36 & 158.41 & 699.06 & 0 & 64.86 & 676.79 & 671.18 & 283.13 & 423.88 & 787.53 & 372.37 & 386.56 & 992.61 & 505.71 & 947.72 & 610.45 & 582.73 \\
911.25 & 706.68 & 745.42 & 570.36 & 427.73 & 511.36 & 621.43 & 829.36 & 413.84 & 590.72 & 69.44 & 667.39 & 613.89 & 758.19 & 195.26 & 715.12 & 64.86 & 0 & 612.00 & 719.16 & 232.54 & 484.71 & 797.75 & 368.50 & 333.97 & 1005.45 & 544.16 & 938.97 & 671.87 & 644.77 \\
1199.97 & 347.07 & 1132.37 & 76.91 & 398.61 & 920.33 & 17.34 & 445.41 & 379.09 & 1020.29 & 662.85 & 1140.17 & 15.61 & 1138.06 & 763.95 & 1047.39 & 676.79 & 612.00 & 0 & 1237.41 & 481.60 & 1077.21 & 1074.43 & 737.92 & 394.94 & 1266.86 & 1056.13 & 1039.97 & 1272.79 & 1239.61 \\
514.39 & 1146.82 & 303.25 & 1223.39 & 875.15 & 349.99 & 1240.14 & 1228.67 & 884.64 & 1056.52 & 649.75 & 137.77 & 1232.84 & 317.39 & 737.31 & 401.86 & 671.18 & 719.16 & 1237.41 & 0 & 951.46 & 322.00 & 499.62 & 992.84 & 846.00 & 600.26 & 911.57 & 791.77 & 678.75 & 296.85 \\
1116.29 & 687.40 & 966.23 & 419.78 & 475.12 & 730.03 & 495.51 & 813.46 & 452.21 & 555.04 & 301.97 & 898.41 & 488.80 & 978.01 & 301.98 & 924.40 & 283.13 & 232.54 & 481.60 & 951.46 & 0 & 702.88 & 996.92 & 274.50 & 381.13 & 1206.58 & 575.16 & 1096.63 & 809.36 & 857.92 \\
754.68 & 1078.33 & 532.00 & 1045.50 & 781.60 & 409.24 & 1084.00 & 1183.32 & 781.70 & 734.62 & 421.83 & 363.96 & 1076.39 & 548.67 & 442.99 & 583.09 & 423.88 & 484.71 & 1077.21 & 322.00 & 702.88 & 0 & 692.81 & 690.72 & 719.87 & 851.71 & 590.32 & 952.81 & 405.70 & 162.42 \\
126.27 & 855.14 & 197.70 & 1092.37 & 676.83 & 312.29 & 1070.04 & 894.75 & 698.26 & 1326.56 & 742.96 & 369.94 & 1064.22 & 185.53 & 925.43 & 109.75 & 787.53 & 797.75 & 1074.43 & 499.62 & 996.92 & 692.81 & 0 & 1159.80 & 707.48 & 209.70 & 1218.06 & 293.77 & 1098.47 & 758.65 \\
1268.61 & 961.78 & 1086.25 & 668.52 & 735.52 & 860.85 & 753.30 & 1087.90 & 714.57 & 283.29 & 417.10 & 977.34 & 747.23 & 1100.51 & 256.46 & 1071.01 & 372.37 & 368.50 & 737.92 & 992.84 & 274.50 & 690.72 & 1159.80 & 0 & 638.24 & 1364.93 & 337.39 & 1306.76 & 649.50 & 818.71 \\
833.58 & 377.01 & 742.65 & 394.51 & 97.92 & 525.69 & 396.02 & 497.06 & 80.32 & 899.87 & 351.30 & 745.47 & 388.92 & 749.47 & 527.99 & 665.35 & 386.56 & 333.97 & 394.94 & 846.00 & 381.13 & 719.87 & 707.48 & 638.24 & 0 & 910.97 & 873.69 & 742.30 & 990.72 & 879.43 \\
99.24 & 1021.20 & 319.80 & 1289.70 & 873.01 & 507.68 & 1261.25 & 1042.86 & 895.69 & 1520.94 & 949.03 & 495.84 & 1255.82 & 303.13 & 1125.97 & 295.25 & 992.61 & 1005.45 & 1266.86 & 600.26 & 1206.58 & 851.71 & 209.70 & 1364.93 & 910.97 & 0 & 1404.86 & 326.74 & 1252.12 & 888.05 \\
1305.62 & 1234.66 & 1093.94 & 991.39 & 969.79 & 905.96 & 1070.37 & 1360.18 & 954.01 & 162.73 & 552.31 & 944.06 & 1063.75 & 1110.15 & 349.55 & 1114.35 & 505.71 & 544.16 & 1056.13 & 911.57 & 575.16 & 590.32 & 1218.06 & 337.39 & 873.69 & 1404.86 & 0 & 1425.06 & 365.95 & 660.70 \\
335.55 & 748.66 & 491.37 & 1077.59 & 676.20 & 545.72 & 1031.13 & 747.29 & 701.67 & 1512.71 & 899.41 & 658.96 & 1026.90 & 479.28 & 1101.01 & 391.48 & 947.72 & 938.97 & 1039.97 & 791.77 & 1096.63 & 952.81 & 293.77 & 1306.76 & 742.30 & 326.74 & 1425.06 & 0 & 1355.23 & 1038.04 \\
1156.57 & 1367.63 & 932.75 & 1220.84 & 1073.98 & 809.59 & 1284.05 & 1485.46 & 1066.09 & 526.33 & 640.12 & 757.48 & 1276.70 & 949.36 & 510.49 & 988.74 & 610.45 & 671.87 & 1272.79 & 678.75 & 809.36 & 405.70 & 1098.47 & 649.50 & 990.72 & 1252.12 & 365.95 & 1355.23 & 0 & 384.99 \\
796.82 & 1232.60 & 574.92 & 1207.74 & 937.61 & 514.19 & 1246.36 & 1334.21 & 939.00 & 817.37 & 583.23 & 394.57 & 1238.75 & 590.93 & 583.42 & 651.03 & 582.73 & 644.77 & 1239.61 & 296.85 & 857.92 & 162.42 & 758.65 & 818.71 & 879.43 & 888.05 & 660.70 & 1038.04 & 384.99 & 0
\end{bmatrix}$$

Figure A1. Pairwise distance matrix.

References

1. Statista. Number of Active Twitter Users. Available online: https://www.statista.com (accessed on 22 December 2022).
2. Abbasi, M.A.; Chai, S.K.; Liu, H.; Sagoo, K. Real-world behavior analysis through a social media lens. In Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, College Park, MD, USA, 3–5 April 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 18–26.
3. Hasan, M.; Orgun, M.A.; Schwitter, R. Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. *Inf. Process. Manag.* **2019**, *56*, 1146–1165. [CrossRef]
4. Abdelhaq, H.; Sengstock, C.; Gertz, M. Eventweet: Online localized event detection from twitter. *Proc. VLDB Endow.* **2013**, *6*, 1326–1329. [CrossRef]
5. Weng, J.; Lee, B.S. Event detection in twitter. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011.
6. Alhumoud, S. Twitter Analysis for Intelligent Transportation. *Comput. J.* **2019**, *62*, 1547–1556. [CrossRef]
7. Hu, B.; Ester, M. Spatial topic modeling in online social media for location recommendation. In Proceedings of the 7th ACM Conference on Recommender Systems, Hong Kong, 12–16 October 2013; pp. 25–32.
8. Rakesh, V.; Reddy, C.K.; Singh, D. Location-specific tweet detection and topic summarization in twitter. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara, ON, Canada, 25–28 August 2013; pp. 1441–1444.
9. Cheng, Z.; Caverlee, J.; Lee, K. You are where you tweet: A content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Niagara Falls, ON, Canada, 25–28 August 2013; pp. 759–768.
10. Ao, J.; Zhang, P.; Cao, Y. Estimating the locations of emergency events from Twitter streams. *Procedia Comput. Sci.* **2014**, *31*, 731–739. [CrossRef]
11. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake shakes Twitter users: Real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 851–860.
12. Imran, M.; Castillo, C.; Diaz, F.; Vieweg, S. Processing social media messages in mass emergency: A survey. *ACM Comput. Surv.* **2015**, *47*, 67. [CrossRef]
13. Graham, M.; Hale, S.A.; Gaffney, D. Where in the world are you? Geolocation and language identification in Twitter. *Prof. Geogr.* **2014**, *66*, 568–578. [CrossRef]
14. Zheng, X.; Han, J.; Sun, A. A survey of location prediction on twitter. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 1652–1671. [CrossRef]
15. Sloan, L.; Morgan, J. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PLoS ONE* **2015**, *10*, e0142209. [CrossRef] [PubMed]
16. Ritter, A.; Clark, S.; Etzioni, O. Named entity recognition in tweets: An experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–29 July 2011; pp. 1524–1534.
17. Liu, X.; Wei, F.; Zhang, S.; Zhou, M. Named entity recognition for tweets. *ACM Trans. Intell. Syst. Technol.* **2013**, *4*, 3. [CrossRef]

18. Liu, X.; Zhang, S.; Wei, F.; Zhou, M. Recognizing named entities in tweets. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; Volume 1, pp. 359–367.
19. Malmasi, S.; Dras, M. Location mention detection in tweets and microblogs. In Proceedings of the Conference of the Pacific Association for Computational Linguistics, Bali, Indonesia, 19–21 May 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 123–134.
20. Jurgens, D.; Finethy, T.; McCorriston, J.; Xu, Y.T.; Ruths, D. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In Proceedings of the Ninth International AAAI Conference on Web and Social Media, Oxford, UK, 26–29 May 2015.
21. Poulston, A.; Stevenson, M.; Bontcheva, K. Hyperlocal home location identification of twitter profiles. In Proceedings of the 28th ACM Conference on Hypertext and Social Media, Prague, Czech Republic, 4–7 July 2017; pp. 45–54.
22. Mahmud, J.; Nichols, J.; Drews, C. Where is this tweet from? inferring home locations of twitter users. In Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, 4–8 June 2012.
23. Ukkusuri, S.V.; Yang, C. *Transportation Analytics in the Era of Big Data*; Springer: Berlin/Heidelberg, Germany, 2019.
24. Chang, H.w.; Lee, D.; Eltaher, M.; Lee, J. @ Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), Istanbul, Turkey, 26–29 August 2012; pp. 111–118.
25. Eisenstein, J.; O'Connor, B.; Smith, N.A.; Xing, E.P. A latent variable model for geographic lexical variation. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, USA, 9–11 October 2010; pp. 1277–1287.
26. Mahmud, J.; Nichols, J.; Drews, C. Home location identification of twitter users. *ACM Trans. Intell. Syst. Technol.* **2014**, *5*, 47. [CrossRef]
27. Flatow, D.; Naaman, M.; Xie, K.E.; Volkovich, Y.; Kanza, Y. On the accuracy of hyper-local geotagging of social media content. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China, 2–6 February 2015; pp. 127–136.
28. Wing, B.P.; Baldridge, J. Simple supervised document geolocation with geodesic grids. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; Volume 1, pp. 955–964.
29. Roller, S.; Speriosu, M.; Rallapalli, S.; Wing, B.; Baldridge, J. Supervised text-based geolocation using language models on an adaptive grid. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju, Republic of Korea, 12–14 July 2012; pp. 1500–1510.
30. Kinsella, S.; Murdock, V.; O'Hare, N. I'm eating a sandwich in Glasgow: Modeling locations with tweets. In Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents, Glasgow, UK, 28 October 2011; pp. 61–68.
31. Paraskevopoulos, P.; Palpanas, T. Fine-grained geolocalisation of non-geotagged tweets. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Paris, France, 25–28 August 2015; pp. 105–112.
32. Hulden, M.; Silfverberg, M.; Francom, J. Kernel density estimation for text-based geolocation. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
33. Mostafa, A.; Gad, W.; Abdelkader, T.; Badr, N. Pre-HLSA: Predicting home location for Twitter users based on sentimental analysis. *Ain Shams Eng. J.* **2022**, *13*, 101501. [CrossRef]
34. Mahajan, R.; Mansotra, V. Predicting geolocation of tweets: Using combination of CNN and BiLSTM. *Data Sci. Eng.* **2021**, *6*, 402–410. [CrossRef] [PubMed]
35. Kleinbaum, D.G.; Dietz, K.; Gail, M.; Klein, M.; Klein, M. *Logistic Regression*; Springer: Berlin/Heidelberg, Germany, 2002.
36. Kibriya, A.M.; Frank, E.; Pfahringer, B.; Holmes, G. Multinomial naive bayes for text categorization revisited. In Proceedings of the Australasian Joint Conference on Artificial Intelligence, Cairns, Australia, 4–6 December 2004; Springer: Berlin/Heidelberg, Germany; pp. 488–499.
37. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef] [PubMed]
38. Wang, L. *Support Vector Machines: Theory and Applications*; Springer Science & Business Media: New York, NY, USA, 2005; Volume 177.
39. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. [CrossRef]
40. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]