Ain Shams Engineering Journal xxx (xxxx) xxx

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

Electrical Engineering

Pre-HLSA: Predicting home location for Twitter users based on sentimental analysis

Aml Mostafa, Walaa Gad *, Tamer Abdelkader, Nagwa Badr

Faculty of Computer and Information Sciences, Ainshams University, Cairo, Egypt

ARTICLE INFO

Article history:
Received 20 February 2021
Revised 16 April 2021
Accepted 22 May 2021
Available online xxxx

Keywords:
Prediction
Home location
Classification
Sentiment analysis
WordNet

ABSTRACT

Identifying the home location of Twitter users is very important in many business community applications. Therefore, many approaches have been developed to automatically geolocate Twitter users using their tweets. In this paper, a new model to predict home location for Twitter users based on sentiment analysis (Pre-HLSA) is proposed. It predicts the users' home location using only their tweets, by analyzing some of the tweet's features. Achieving this goal allows providing geospatial services, especially in the epidemic dispersion. The Pre-HLSA represents user tweets as a set of extracted features and predicts the users' home locations by analyzing their tweets to find sentiments and polarities, even in the absence of geospatial clues. Then, different classifiers are applied. The experimental results show a promising performance compared to the previous methods in terms of accuracy, mean and median performance measures. It achieves up to 85% accuracy, 223 km mean, and 96 km median.

1. Introduction

Recently, microblogging services such as Twitter have been in rapid growth More than 300 million active users write more than 500 million tweets every day [1]. Many studies have been spawned to take advantage of these human-powered networks and build applications such as advertisements [2], restaurant detection, crime prediction [3], epidemic inspiration, etc. Such applications benefit from information about users' location.

However, data about Twitter user's location are very important, users are very slow to write their profiles and complete their geolocation features. Only 3% [4,5] of Twitter users are interested in geotagged features as granular, therefore location data are very sparse. Moreover, users may write in their profiles inaccurate home location and inaccurate time zone information because of security, and no restrictions are applied [6,7]. In 2009, Twitter supports an additional property which is per-tweet geo-tagging. This property tracks Twitter users by associating latitude and longitude for each tweet.

In this paper, predicting home location for Twitter users based on the sentiment analysis (Pre-HLSA) model is proposed. The proposed model overcomes the problem of location sparseness by using only the tweet content to predict the user's location. It is based on the tweet content (text), no need for user IP information or users logging information. The pre-HLSA predicts the home location by analyzing users tweets to extract the most important features, sentiment analysis methods are applied to identify user's polarity, then nine classifiers are applied.

The proposed predicting home location for Twitter users based on sentiment analysis consists of

- Pre-processing.
- Feature Extraction.
- Sentimental analysis.
- Classification.

In the first step, tweets are pre-processed to remove unwanted web site addresses in texts and clear irrelevant spaces and marks. In the feature extraction step, the most important features are extracted such as userID, data, time. The proposed Pre-HLSA adopts per-tweet geo-tagging properties to extracts coordinate 1 and coordinate 2 as important features. Distance feature is calculated. Tweet text is represented as numeric values using a Term

* Corresponding author.
E-mail addresses: sofha@gmail.com, Walaagad@cis.asu.edu.eg (W. Gad).

Peer review under responsibility of Ain Shams University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.asej.2021.05.015
2090-4479/© 2021 THE AUTHORS. Published by Elsevier BV on behalf of Faculty of Engineering, Ain Shams University.

frequency-inverse document. Time feature is categorized into four intervals: 00 to 07, 07 to 12, 12 to 18, and 18 to 24. The date is represented as seven categories: Monday to Sunday.

Then, sentiment analysis is applied to identify the user sentiments. Each tweet text is analyzed to identify user polarity. Sentimental analysis output can be negative, positive, neutralize, and compound based tweet context. Finally, nine classifiers, Linear Regression, linear_model SGD Regressor, linear_model Theil Sen Regressor, Support Vector Machine (SVM), Decision Trees, Random Forest, Neural Networks, K-nearest neighbour (KNN), Gradient Boosting Regressor, are applied.

The proposed predicting home location for Twitter users based on sentiment analysis (Pre-HLSA) model is evaluated using the dataset [8]. Three experiments are done. The first experiment is the full data (FD). In experiment two, the Pre-HLSA is evaluated using the non-zero distance feature (NZD). Experiment three is done using non-zero distance and non-zero compound (NZD + NZC). Three performance measures are used to assist Pre-HLSA: accuracy, mean, median. The best performance is based on a decision tree. It records 85% accuracy, 223 km mean, 96 km median.

The paper is organized as follows. Section 2 presents a literature review. Section 3 explains the predicting home location for Twitter users based on the sentiment analysis (Pre-HLSA) model. Section 4 presents evaluation methods and performance measures. Finally, the conclusion is in the last section.

2. Related work

Recently, there have been an increasing interest in the geolocation prediction problem, and many research works proposed their solutions [8–16]. These solutions can be classified into four main approaches: the tweet content, the network or relationship (friends and followers), the context of the tweet, and finally a hybrid mix of two or more factors together. The work in [8] is one of the earliest in this field. The authors used linguistics for applying topic models to predict the home location. For instance, lakers and iTunes may relate to "popular music" topic. The term "pearl" may be also related to the "popular music" topic for the Boston region. It was a similar topic, but each word had a different frequency. In [9], the authors introduced a Sparse Additive Generative (SAGE) model in the topic model. In [8] and [9], authors applied their models to predict the home location of the user. Furthermore, using the Streaming Latent Dirichlet Allocation (SLDA) model, they applied the topic model and used the features of the text and context for the prediction process. Moreover, they applied the trends of the topic model and topical location. If some spots have the same name, they lead to solving the problem of the geographical disambiguation.

In [10], authors utilized a grid cell representation of locations which has the smallest KL difference. They evaluated a language model that eliminates the bag-of-words presumption. In [11], the authors got the propagation of the label that would be influenced by highly connected nodes (i.e., celebrities with many followers) and found the nodes that are not connected to any other nodes that are gathered. Moreover, they recognized the celebrities' nodes by the number of mentions based on a graph relationship. On the other hand, for the nodes that has no labelled relationship neighbours in the graph, they estimated the relationship using a proposed model, that is used in [12], called content-based method. Authors in [13] utilized a multilayer perceptron (MLP) to detect the home location of the user. They assumed a normalization of a bag-of-words description of the content, to obtain a region discretized by applying k-means or k-d tree. Deep learning techniques play a role in feature extraction automatically.

In [13] and [14], authors used Neural Network models to solve and analyze the problems of geolocation in Twitter. In [15], the authors applied a classification model to solve local word association problems by using the Spatial Variation model. In [16] the authors applied the same method (The Spatial Variation model) to a Korean tweet dataset. In [17], the Neural Network model is also applied to classify the most geotagged location of the user as a center of the classifier, rather than calculating the median of the geometric. In addition, the authors reconstructed the classification process to calculate the center area with more accurate cells till they obtain a smaller area than a predefined area. They considered the final center area as the home location of the user.

The authors in [18] applied the Spatial Variation model, they supposed that every word has a geographical centre, a centre has a frequency, and a distribution ratio, by calculating the probability of the word in a location among distance which is the center of the proportional. Taking into consideration the simple words, the model applies a one-peak distribution by fitting this model, the words become features. They labelled 19,178 words in a dictionary as either local or non-local. Then they apply a classification methodology to all other words in the tweet database.

Table 1 shows a summary for the previous studies. Four main factors are used for predicting home location:

- Content, which depends on tweet text.
- Network, which is based on relationships such as friends and followers.
- Time, which is the posting time.
- Hybrid, which combines between two or more of the previous factors together.

Moreover, a comparison is done based on the granularity level. The granularity level is categorized as

- Administrative level, such as country, city, or state where the users stay in.
- Geographical grid, in which the ground is partitioned into cells, and the home location is represented by the cells they stay in.
- Geographical coordinates, which are represented using the longitudes and the latitudes.

3. Predicting home location for Twitter users based on sentiment analysis model

Most Twitter users are not interested in granular information, nor to fill information about their location. Users location information is very important in many applications, such as event prediction, and services recommendation. In this paper, predicting home location for Twitter users based on sentiment analysis (Pre-HLSA)

Table 1
A summary showing home location prediction (HLP) studies.

Study Reference	Solution Approach	Granularity	Performance Measures
[8]	Content	State	Mean, Median
[9]	Content	Coordinates	Mean, Median
[10]	Content	Grid	Mean, Median
[11]	Hybrid (Content + network)	Coordinates	Acc, Mean, Median
[12]	Hybrid (Content + network)	Coordinates	Acc, Mean, Median
[13]	Content	Grid	Acc, Mean, Median
[14]	Hybrid (Content + network)	City	Acc, Mean, Median
[15]	Content	City	Acc, Mean,
[16]	Content	Coordinates	Mean

model is proposed. The Pre-HLSA uses sentiment analysis, and classification to predict user's location, which helps to overcome the sparseness of location problems, as shown in Fig. 1.

3.1. Pre-processing and feature extraction

In this step, the tweet text is pre-processed and tokenized to remove the irrelevant tokens in the text, such as web site links, hashtags, stop words, etc. The following rules are applied to the tweet text to remove unwanted words:

- Remove punctuation, such as: . , ? ! \ ’ “ ( ) : ; .
- Convert repetitions of 3 or more letters to 2 letters, such as boooook --> book.
- Check the validity of each word.
- Handle emoji such as (,): , :’(,)’: ,) :, :(:-D, x-D , XD , X-D
- Convert all letters to lower case.
- Convert URLs to the word URL.

An example for pre-processing step is shown as follows:

Tweet before pre-processing: “Success is just a war of attrition. Sure, there is an element of talent you @USER_4e7f65fc. But if you just stick around long enough, eventually something is going to happpen :).”

Tweet after pre-processing: “success is just a war of attrition sure there is an element of talent you but if you just stick around long enough eventually something is going to happen EMO_POS”

After, removing irrelevant tokens, term frequency-inverse document frequency (TF-IDF) Equation 3 is applied. Each tweet text is represented as a vector of words. The word weigh is represented by TF-IDF. Term frequency (TF) is defined as number of occurrences of a word *(w_i). efined as inverse document frequency as shown in*

$$tf(w_i) = \frac{f(w_i, t_j)}{numberofwordint_j} \tag{1}$$

where f*($w_{i,}$ t_j))* is the frequency of *(w_i)* in tweet *j* .

$$idf(w\_i, T) = LogN/(t \in T : w\_i \in t\_j) \tag{2}$$

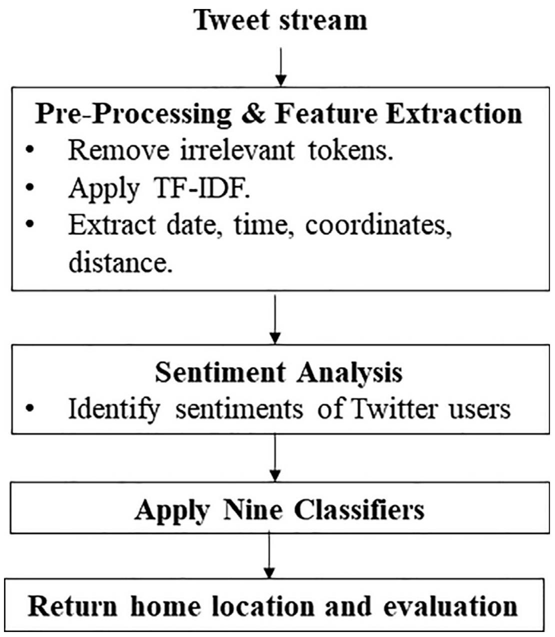
Tweet stream
Pre-Processing & Feature Extraction
• Remove irrelevant tokens.
• Apply TF-IDF.
• Extract date, time, coordinates, distance.
Sentiment Analysis
• Identify sentiments of Twitter users
Apply Nine Classifiers
Return home location and evaluation

Fig. 1. The framework of predicting home location for Twitter users based on sentiment analysis (Pre-HLSA) model.

where *T* is Twitter corpus. *N* is the number of tweets in corpus, *N* =|*T*|.

$$TF - IDF(w\_i)) = tf(w\_i).idf(w\_i, T) \tag{3}$$

The proposed Pre-HLSA extracts some important features, such as userID, date, time, cooridnate1(latitude) and coordinate 2 (longitude) features to be used in the training step. Coordinate 1 and Coordinate 2 features are used to calculate the distance feature between tweets of the same user. The first tweet of each user is considered as the baseline. the home location of the user. Time feature is categorized to four intervals: 00 to 07, 07 to 12, 12 to 18, and 18 to 24. The date is represented as seven categories: Monday to Sunday.

3.2. Sentimental analysis

In this step, WordNet and sentiment analysis are applied for tweets to identify and extract their sentiments and polarity. WordNet is a lexical network, which is used for semantic meaning relations. It handles the words as nodes in a network and handles the edges of the network as relations of the synonyms. In WordNet, the word meaning is called a word-sense. Words are originated from their root words. For instance, “love” is lessened to “adore”. It includes forceful rules of grammar. In this step, Emotion Artificial Intelligence (AI) which is a part of Natural Language Processing (NLP) which is used to categorize the tweet text into one of the following labels: positive, negative, neutral, or compound.

The Pre_HLSA adapts to the Valence Aware Dictionary for sEntiment Reasoner (VADER) [19]. It aims to analyze tweet content to obtain users opinion and automatically classify their sentiment polarity. VADER is a lexicon and rule-based method to get the users opinion sentiment. The sentiment polarity is categorized into four types: positive(pos), negative(neg), neutral(neu), and compound (comp). For each tweet, there is a given score for each type as shown in Table 2. The compound record is a metric that counts the summation of all the lexicon extreme ratings which calculates the lexicon normalized between −1(the extreme of negative) and + 1 (the extreme of positive). A threshold is predefined to differentiate between positive, negative, and neutral sentiments. A Tweet has a positive sentiment if the compound value is greater than threshold, and it has a negative sentiment if the compound value is less than the threshold. Otherwise, a tweet is considered neutral sentiment. The overall sentiment score for the tweets has a rate ranging from −2 (extremely negative) to + 2 (extremely positive).

3.3. Classification

In this step, classification is applied using nine machine~~s~~ learning classifiers as follows.

Linear Regression is one of the most popular Machine Learning algorithms that are used in prediction task. It is based on a given independent variable (x) to predict and get the output variable (y). The Linear Regression is defined as follows.

$$y = a + bx \tag{4}$$

where y is the predicted value, b is the rate of predicted scores, x is the independent variable and a is the intersect level of y - when x = 0.

Theil-Sen estimator is a very simple Regressor used for prediction tasks. The concept of the simple predictor is that if the data contains N pairs of (a, b, c) values, calculate all the slopes between pairs of points and determine the median as the estimation of the regression slope. Using that slope, move a line through each pair of (a, b, c) values to obtain N intercepts. Choose the median of the intercepts as the estimation of the regression as follows.

Table 2
an example of sentiment scores.

Text	Pos	Neg	Neu	Compound
yes, just got here	0.423	0.0	0.577	0.296

$$y = m_{median} - nx_{median} \tag{5}$$

where y is the predicted intercept, m_{median} is the median of the response variables, n is the estimated slope and x_{median} is the median of the analytical variables.

Stochastic Gradient Descent (SGD) Regressor; according to the meaning of the word ‘Stochastic’, it means randomly that a method combined with a random probability. Therefore, in Stochastic Gradient Descent, instead of the full data set for each iteration, a few specimens are selected randomly. Gradient Descent is a common optimization method in Machine Learning. A gradient is essentially the slope of a function. Mathematically, it can be explained as the partial derivatives of a set of parameters concerning its inputs. Thus, the more the gradient increases, the more precipitous the slope is. Stochastic estimator examines the minimizing problem as

$$y(x) = \frac{1}{n}\sum_{i=1}^{n} y_i(x) \tag{6}$$

where *y(x)* is the predicted value, n is the number of observations, i is the iteration, and $\times$ *is* the given input.

Support Vector Machine (SVM) considers any data item that being a point in n-dimensional (where n is the number of features that we used in each phase that we mentioned before), the power and effect of SVM are that the Support Vector Machine (SVM) can be reformulated and combined to the internal product of any two given observations, rather than the observations themselves.

A Decision Tree is a classifier using tree-structured in the classification process where the features are the nodes, the decision rules are the branches, and the outcome of the classifier is the leaf node. Information gain is calculated as

$$Gain(A) = information(T) - entorpy(A) \tag{7}$$

where A is all possible attributes values.

$$information = -\sum_{i=1}^{c} Pi\log_2(pi) \tag{8}$$

$$Entorpy = Pi\log_2(pi) \tag{9}$$

where c is the number of classes, pi is the corresponding frequency of the observed class, *entropy* is the measure of a random variable Entropy uses the probability of a confident result to decide how the node should branch.

The Random Forest is a simple classifier model that is mostly used. It is an aggregate of randomized decision trees. The Random Forest is a model of classification consisting of multiple decision trees. The tree consists of many branches and each branch is a decision, the random forest chooses the final decision based on the majority of the decision tree.

$$Gini = 1 - \sum_{1=1}^{c} (Pi)^2 \tag{10}$$

The *Gini* coefficient is a value number pointed to measuring the degree of variation in a distribution. where pi is the corresponding frequency of the observed class, c is the number of classes.

Neural Networks (NN) are artificial methods that extracted from biological Neural Networks, that can be used as a Machine Learning method that depends on training data to understand and get the logical results of these data sets. Neural networks are a combination of computational and mathematical models. This achievement has been directed to improvements in infinite automation. It depends on the weight of the connection between neurons and each other. The greater of this value, the large number of the weight, the bias is a hidden layer, each neuron is not the input layer has a bias connected to it, and the bias has a value as the weight.

$$f\left(b + \sum_{i=1}^{n} x_i w_i\right) \tag{11}$$

where b is the bias, i is a number of the iteration, n is the number of all input, x is the neuron input and w is the weight of the neuron.

K-Nearest Neighbours (KNN) is one of the most essential classification algorithms in Machine Learning. The KNN algorithm believes in similarity that the main idea in KNN is that similar things exist nearby. KNN calculates the distance between points and each other on a graph. Euclidean distance is calculated as follows.

$$D = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{12}$$

where D is the Euclidean distance, i is the number of iteration and n is the number of all inputs, $x_i - y_i$ are Euclidean vectors starting from the origin point.

Gradient Boosting Aggregate is a set of weak classifier models to obtain a strong model that can be used in the prediction task such as decision tree, so the aggregation in both models decline the gradient of a loss function.

4. Result and analysis

The proposed predicting home location for Twitter users based on sentiment analysis (Pre-HLSA) model is evaluated using the dataset published in [8]. It was collected from Twitter’s streaming Application Programming Interface (API) in March 2010, which is a 15% description of all daily messages. In [8], the authors included only geotagged data within the bordering of The United States. Furthermore, they cleared the left data to hold only users followed by and following less than 1,000 people to avoid celebrities. The final dataset used consists of about 380,000 tweets and 9,500 users. In the experiments, the data is divided into 75% for training, and 25% for testing.

After Pre-processing, feature extraction and sentimental analysis are done. Tweets are represented as set of records. Each record consists of features such as userID, date, time, coordinate 1, coordinate 2, distance, tweet content (weighted tokens), sentiment polarity (positive, negative, neutral, compound). The features have numeric values, and ready for the training process. Dataset is divided three sub-datasets for three different experiments. Experiment 1, Full data (FD): all the data that is ready for training. Experiment 2, No zero distance (NZD): all the data except the distances equal to zero. Experiment 3, No zero distance and no zero compound (NZD + NZC): all the data except the distances equal to zero and compound equal to zero. Fig. 2 shows the effectiveness of negative, positive, and neutral features in the training step.

4.1. Performance measures

Three performance measures are used to evaluate the predicting home location for Twitter users based on the sentiment analysis (Pre-HLSA) model: accuracy within 161 km, mean and median in km. Taking into consideration that The higher numbers are more beneficial for Acc@161. The few numbers are better results for mean and median as it shows the error.

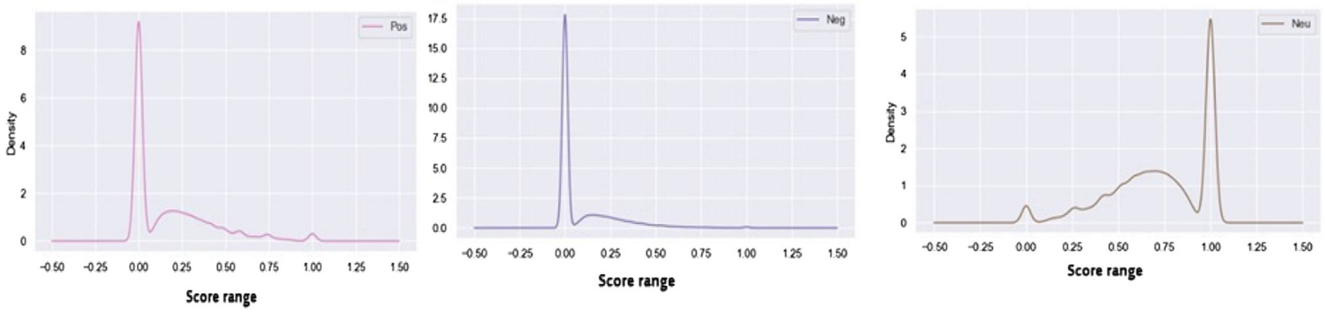

Fig. 2. The effectiveness of Neutral, Negatives and Positive features.

According to [15], the authors show that 30% of users are located within 10 miles of their true location, and 51% of users are located within 100 miles after investigating a range from 0 to 4,000 miles. Pre-HLSA defines the accuracy at 100 miles (161 km) as follows.

$$Acc@161\left(d,\bar{d}\right) = \frac{1}{n_{users}} \sum_{i=0}^{n_{users}-1} 1\left(errordistnceu_i < 161km\right) \tag{13}$$

where ***d*** is the real distance, that is labeled in the dataset, and ***d'*** is the error distance calculated from the proposed model of the user (***u***), ***n*** is the number of the users, and $errordistnceu_i isthegreatcircle\{\mathrm{d},\bar{\mathrm{d}}\}$.

The second performance measure is the Simple Accuracy Error (SAE) metric which is the distance from the predicted home location to the correct current home location of the user that is represented by mean and median.

$$mean = \frac{1}{n_{users}} \sum_{i=0}^{n_{users}-1} \{ErrorDistance(ui)\} \tag{14}$$

$$median = median_{i=0}^{n_{users}-1}\{ErrorDistance(ui)\} \tag{15}$$

where ***d*** is the real distance, that is labelled from the data and ***d'*** is the error distance calculated from the proposed model of the user (u), n is the number of the users.

4.2. Results and analysis

The proposed Pre-HLSA is evaluated using three experiments: FD, NZD, and NZD + NZC, and nine different classifiers as shown in Table 3. The best accuarcy@161 is based on decision tree (DT) classifier. It records 45% for first experiment, 73% for second, and 85% for the third one. The accuracy performance measure is enhanced in experiment one and experiment two, because the sparse records are not included. Table 3 shows a comparison between experiments and using different classifiers in terms of accuarcy, mean and median performance measures. The Pre-HLSA records 495 mean and 198 median for full data experiment, 318 mean and 105 median for non-zero distance experiment (NZD), and 223 mean and 96 median for the non-zero distance and non-zero compound experiment.

Fig. 3 summarizes the proposed Pre-HLSA performance in terms of accuracy, mean and median using the nine classifiers: Linear Regression (LR), linear_model SGD Regressor (SGD), linear_model Theil Sen Regressor (Theil), Support Vector machine (SVM), Decision Trees (DT), Random Forest (RF), Neural Networks (NN), K-nearest neighbour (KNN), Gradient Boosting Regressor (Boosting). Table 4 shows the result of the proposed model which is compared to different studies that use the same database.

4.3. Results interpretation and discussion

The outperformance of the proposed model attributes to its three steps: pre-processing, sentiment analysis and classification. In the pre-processing, words are stemmed, and repetitive words are removed. Then, frequent words are extracted using term frequency-inverse document frequency (TF-IDF). The frequent term is identified by a predefined threshold. By extracting the most important words, the proposed model considers only a subset of words that has a geographical reference to a specific location compared to the other words in the dataset. As shown in Experiment 1, Pre-HLSA, using the decision tree classifier, achieved 45%, 495 km, and 198 km compared to the results of Rahimi et al. (2017) [13], which records 39%, 865 km and 412 km in accuracy, mean, and median, respectively. Roller et al. (2012) [20] use adaptive grid schemes based on k-d trees and uniform partitioning which records 35.9%, 897 km, and 432 km in accuracy, mean and median. Cha et al. (2015) [21] is an unsupervised representational which is based on sparse coding and dictionary learning. They report 581 km mean and 425 km median. Hulden et al. (2015) [22] use kernel- density as a geodesic grid and allows for the use of much smaller grids with less data. Mean and median are recorded to 765 km and 357 km. We claim that this work is pioneer in using sentiment analysis to predict home location for Twitter users. The home location prediction solutions are based on tweet text, tweet users' network or both. Table 5 shows a comparison of the solution methods used in the different proposals.

In the proposed Pre-HLSA, sentiment analysis is conducted to determine the sentiment attitude of the tweet user with respect to the tweet location. There is a strong correlation between user emotions and its locations. The polarity analysis enables the proposed model to quantify the sentiment of the text, and then categorizes the tweet text into positive, negative, or neutral. In Experiment 3, Pre-HLSA achieves 85%, 223 km and 96 km in terms of accuracy, mean and median. Nine classifiers are used to evaluate the proposed model. The dataset is divided into 80% for training and 20% for testing. Results show that the decision tree outperforms the other classifiers. The output of decision tree is much consistent with the objective of the Pre-HLSA. Decision tree determines the best predictors for the output, and outputs better results when there are strong relationships between input and output. Therefore, the decision tree classifier proves that there is a strong relationship between the extracted features: tweet words weights (frequency) and user attitude (emotions), which serves to predict user's location.

5. Conclusion

In this paper, a new prediction home location for Twitter users based on sentiment analysis (Pre-HLSA) is proposed. Pre-HLSA pre-

Table3
The performance of Pre-HLSA in terms of accuracy@161, Mean and Median.

Classifier	**FD**			**NZD**			**NZD + NZC**		
	Acc	Mean	Median	**Acc**	Mean	Median	**Acc**	Mean	Median
Linear Regression (LR)	25	624	203	59	458	192	70	298	102
SGD Regressor (SGD)	15	732	315	50	613	287	62	572	225
Theil Sen Regressor (Theil)	41	527	306	63	378	129	72	260	103
SVM	35	620	325	68	436	241	81	278	133
Decision Tree (DT)	**45**	**495**	**198**	**73**	**318**	**105**	**85**	**223**	**96**
Random Forest (RF)	32	760	376	60	672	268	72	504	137
Neural Networks (NN)	23	844	382	68	632	253	83	509	157
KNN	24	818	328	56	529	228	77	429	164
Gradient Boosting (GBR)	20	554	327	62	489	213	73	380	164

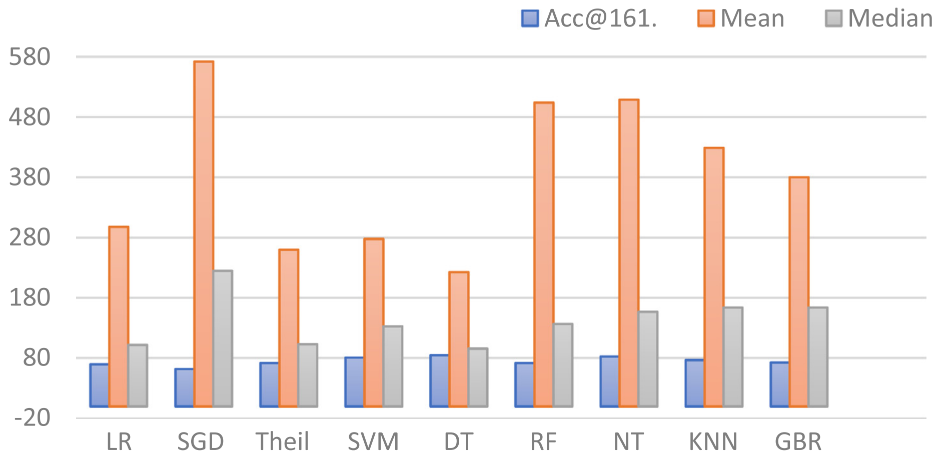

Fig. 3. The performance of of Pre-HLSA in terms of Acc@161, Mean and median in km.

Table 4
Performance of the proposed Pre-HLSA compared to other studies.

Study	**Acc@161**	**Mean**	**median**
Eisenstein et al. (2010)	–	900	494
Eisenstein et al. (2011)	–	845	501
Roller et al. (2012)	35.9	897	432
Rahimi et al.)2015(	38	880	397
Cha et al. (2015)	–	581	425
Hulden et al. (2015)	–	765	357
Rahimi et al. (2017)	39	865	412
Proposed Pre-HLSA	**85**	**223**	**96**

Table 5
Comparison between Pre-HLSA and the other home prediction proposals.

Proposal	Solution Method
Roller et al. (2012)	Adaptive grid schemes based on k-d trees and uniform partitioning. (text- based location)
Cha et al. (2015)	Unsupervised representational which is based on sparse coding and dictionary learning. (text-based location)
Hulden et al. (2015)	Based on a kernel- density to alleviate some of the sparse data problems associated with geolocating documents on a discretized surface modelled as a geodesic grid and allows for the use of much smaller grids with less data. (text-based location)
Rahimi et al. (2017)	Incorporates text-based model into a network-based model and use model embeddings to extract local terms. (text-based + network based)
Pre-HLSA	Sentiment analysis is conducted to determine the sentiment attitude of the tweet user with respect to the tweet location (text-based + sentiment based).

dicts the uses home location based on the tweets content, sentiment analysis and classifications. In the Pre-process step, all unwanted tokens are removed, and tweet tokens are assigned weights. After that, feature extraction is done to extract the important feature such as time, date and calculate distance feature. Sentiment analysis method is applied to identify the Twitter user's sentiment polarity as important extracted feature. Finally, nine classifiers are applied to predict home location. Three experiments are conducted to evaluate the proposed model. Three performance measure are used to assist proposed model. The experimental results show a promising performance compared to other studies. Pre-HLSA records 85, 223, 96 in terms of accuracy, mean and median performance measures respectively.

Declaration of Competing Interest

None.

References

[1] Ajao Oluwaseun, Hong Jun, Liu Weiru. A survey of location inference techniques on Twitter. Journal of Information Science 2015;41(6):855–64.

[2] Shen-Shyang Ho, Mike Lieberman, Pu Wang, and Hanan Samet, Mining future spatiotemporal events and their sentiment from online news articles for location-aware recommendation system, In Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems. Redondo Beach, California, pages 25–32.

[3] Mingjun Wang and Matthew S. Gerber, Using Twitter for Next-Place Prediction, with an Application to Crime Prediction. IEEE Symposium Series on Computational Intelligence.

[4] Cheng, Z.Caverlee, J., & Lee, K, A content-driven framework for geolocating microblog users, ACM Transactions on Intelligent Systems and Technology, 4 (1) 2:1–2:27.

[5] Graham, M., Hale, S. A., & Gaffney, D, Where in the world are you? geolocation and language identification in twitter, The Professional Geographer, 66(4), 568–578.

[6] Satyen Abrol and Latifur Khan, Tweethood: Agglomerative clustering on fuzzy k-closest friends with variable depth for location mining, In Proceedings of the IEEE 2nd International Conference on Social Computing (SocialCom'10). 153–160.

[7] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi, Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles, In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 237–246.

[8] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, A latent variable model for geographic lexical variation, In EMNLP, pages 1277–1287, 2010.

[9] J. Eisenstein, A. Ahmed, and E. P. Xing, Sparse additive generative models of text, in Proc. 28th Int. Conf. on Machine Learning, 2011, pp. 1041–1048.

[10] B. P. Wing and J. Baldridge, Simple supervised document geolocation with geodesic grids, in Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, 2011, pp. 955–964.

[11] Rahimi A, Cohn T, Baldwin T. Twitter user geolocation using a unified text and network prediction model. In: Proc 53rd Meeting of the Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing of the Asian Federation of Natural Language Processing. p. 630–6.

[12] A. Rahimi, D. Vu, T. Cohn, and T. Baldwin, Exploiting text and network context for geolocation of social media users, in Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 1362–1367.

[13] A. Rahimi, T. Cohn, and T. Baldwin, A neural model for user geolocation and lexical dialectology, in Proc. 55th Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers, 2017, pp. 209–216.

[14] Miura Y, Taniguchi M, Taniguchi T, Ohkuma T. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In: Proc 55th Annual Meeting of the Association for Computational Linguistics. p. 1260–72.

[15] Cheng Z, Caverlee J, Lee K. You are where you tweet: a content-based approach to geo-locating twitter users, in Proc. In: 19th ACM Conf. on Information and Knowledge Management. p. 759–68.

[16] K. Ryoo and S. Moon, Inferring twitter user locations with 10 km accuracy, in Proc. 23rd Int. World Wide Web Conf. Companion Volume, 2014, pp. 643–648.

[17] Z. Most Twitter users are not interested in granular information, nor to fill information about their Cheng, J. Caverlee, H. Barthwal, and V. Bachani, Who is the barbecue king of texas?: a geo-spatial approach to finding local experts on twitter, in Proc. 37th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2014, pp. 335–344.7.

[18] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, Spatial variation in search engine queries, in Proc. 17th Int. Conf. on World Wide Web, 2008.

[19] Shihab Elbagir and Jing Yang, Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment, Proceedings of the International MultiConference of Engineers and Computer Scientists 2019 IMECS 2019, March 13-15, 2019, Hong Kong.

[20] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge, Supervised text-based geolocation using language models on an adaptive grid, in Proc. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language

[21] Miriam Cha, Youngjune Gwon, and H.T. Kung, Twitter geolocation and regional classification via sparse coding, In Proceedings of the 9th International Conference on Weblogs and Social Media (ICWSM 2015), pages 582–585, Oxford, UK.

[22] M. Hulden, M. Silfverberg, and J. Francom, Kernel density estimation for text-based geolocation, in Proc. 29th AAAI Conf. on Artificial Intelligence, 2015, pp. 145–150.

Aml Mostafa is a master student at the information system department, Faculty of Computer and Information science, Ain Shams University, Cairo, Egypt. She received the BSc degree in Computer and Information Sciences in 2014 from Ain Shams University. Her current research interests include data mining, machine learning, semantic analysis.

Walaa Gad is associate professor at the information system department, Faculty of Computer and Information science, Ain Shams University, Cairo, Egypt. She received the BSc and the MSc degrees in computers and information sciences in 2000 and 2005 respectively, from Ain Shams University, Cairo, Egypt. She was a Ph. D student in the Pattern and Machine Intelligence (PAMI) Group, Faculty of Electrical and Computer Engineering, University of Waterloo, Canada. She received her Ph.D in 2010. The work was done jointly between Faculty of Computers and Information Sciences, Ain Shams University and University of Waterloo in Canada. Her current research interests include data science, semantic web and machine learning, data warehouse and data analytics.

Tamer Abdelkader received the B.Sc. degree in electrical and computer engineering and the M.Sc. degree in computer and information sciences from Ain Shams University, Cairo, Egypt, in 2003, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Waterloo, Ontario, ON, Canada, in 2012. After graduation, he was with the University of Waterloo as a Postdoctoral Researcher and a Visiting Researcher. He worked as the Manager of the Information and Technology Research Consultancy Center, Ain Shams University, Cairo, Egypt. He also worked as an Information and Technological Consultant in several governmental and private companies, including the Information and Communication Technology Project, Egypt, and the Ministry of Electricity. He is currently an Associate Professor and Vice-Dean for Community Services and Environmental Affairs with the Faculty of Computer and Information Sciences, Ain Shams University. He is the author of several publications in IEEE Transactions and other ranked journals and conferences. His current research interests include network and information security, delay-tolerant networks, resource allocation in wireless networks, vehicular networks, and energy-efficient protocols.

Nagwa Badr is a professor and dean at Faculty of Computer and Information Sciences ,Ain Shams University. She received the B.Sc. degrees in Computer Science in 1996 and PhD from Liverpool John Moorse University, U.K. in 2003 in Software Engineering and Distributed Systems. She had done postdoctoral studies in Glasgow University, U.K . She is a head of the committee that contributed in research projects funded by national and international grants of Information Systems, Bioinformatics, Business Analytic and Health Informatics (i.e. http://www.heal-plus.eu/).Her current research areas are in Software Engineering, Cloud Computing, Big Data analytics, social networking, Arabic search engines and Bioinformatics.