بسم الله الرحمن الرحيم

∞∞∞

تم رفع هذه الرسالة بواسطة / هناء محمد علي

بقسم التوثيق الإلكتروني بمركز الشبكات وتكنولوجيا المعلومات دون أدنى مسئولية عن محتوى هذه الرسالة.

ملاحظات: ....

Ain Shams University
Faculty of Computer and Information Sciences
Information System Department

Location Prediction Using Data Mining Techniques

Thesis submitted as a partial fulfillment of the requirements for the degree of Master of Science in Computer and Information Sciences

By

Aml Mostafa Ismaiel

Teaching Assistant at Information System Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Under Supervision of

Prof. Dr. Nagwa Badr
Professor and Dean of Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Tamer Abdelkader
Associate Professor and Vice-Dean for
Community Services and Environmental Affairs with the Faculty of Computer and Information Sciences, Ain Shams University

Cairo-Egypt
2022

STATEMENT

This dissertation is submitted to Ain Shams University in partial fulfillment for the degree of Master of Science in computer and Information Sciences.

The work included in this dissertation was carried out by the author at the Information System Department, faculty of computer and information science, Ain Shams University.

No part of this dissertation has been submitted for a degree or a qualification at any other university or institute.

Name: Aml Mostafa Ismaiel.

Signature: *Aml Mostafa.*

Date: 10 /2021.

Acknowledgment

I am extremely thankful to **Prof. Dr. Nagwa Badr**, for allowing me to do and finish this research, she is guiding and helping me to solve all problems that facing me during the research process. she always there any time I need it.

Also, I like to express my thanks and appreciation to **Dr. Tamer. Abdelkader,** for his role to guide me in solving all problems to complete the research (Location Prediction using data mining techniques), I fully appreciate his support, time, and effort.

Also, I like to express my appreciation to **Dr. Walaa Khaled,** for her ideal supervision, recommendations given by her to help me to grow as a scientific researcher. Her effort spent in discussions during the work leads to reaching great results and achieve our goal. She supported me with All forms of support.

Dedication

I want to dedicate this thesis:
To my all supervisors who did not abandon me through all the duration of research and working on this thesis, they always encourage me and support me with all forms of supports.

To all my family who is always there in my side all the time, encouraging, supporting, and helping me with all their efforts and support to reach this stage.

Abstract

The rapid use of social media made location prediction the key to research studies based on-location services such as, advertising, recommendations, climatological forecast, and security system. Locations are the center of information for these applications. According to millions of users who post tweets every day, the geographical coordinates are often hidden in Twitter due to privacy reasons. 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. Depending on the importance of catching the location of the users and the rapid usage of Twitter, Location prediction on Twitter has been a point of research in many studies.

This thesis work provides a comprehensive overview of the prediction of the user's location on Twitter, which focuses on the home location prediction and tweet location prediction. This is achieved by defining the inputs of these two research views that are content, network, and context, and then proposing two new location prediction models.

The First proposed model is to predict the tweet location based on the KNN-Sentimental Analysis (KNNSA) model. Predicting the tweet location based on the KNN-sentiment analysis (KNNSA) extracts text features from the tweet in addition to the date and time features. Then, applying sentimental analysis and classifying the data by K-nearest neighbors (KNN) classifier. The (KNNSA) is evaluated and compared to the previous work and it achieves better performance in terms of root mean squared error (RMSE) and of the mean absolute error (MAE).

The second proposed work is to predict home location for Twitter users based on sentiment analysis (Pre-HLSA). It predicts the users' home location using only their tweets, by analysing 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 analysing their tweets to find sentiments and polarities, even in the absence of geospatial clues. Then, different classifiers are applied by applying sentimental analysis 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.

List of Publications

- "Predicting the tweet location based on KNN-Sentimental Analysis", Published in**:** 2020 15th International Conference on Computer Engineering and Systems (ICCES). 2020. Cairo, Egypt.
- "Pre-HLSA: Predicting Home Location for Twitter Users based on Sentimental Analysis" Published by Elsevier BV on behalf of Faculty of Engineering, Ain Shams University. Cairo, Egypt. https://doi.org/10.1016/j.asej.2021.05.015 . 2021
- "A Location Prediction Methods: state of art", International Journal of Intelligent Computing and Information Science (IJICIS), 2021.Cairo, Egypt.

Contents

List of Figures

List of Tables

List of Abbreviations

TLP	Tweet Location Prediction.
HLP	Home Location Prediction.
POI	Point of Interest.
DBN	Dynamic Bayesian Network.
IR	Information Retrieval.
NN	Neural Network.
ILF	Inverse Location Frequency.
ICF	Inverse City Frequency.
TF-IDF	Term Frequency–Inverse Document Frequency.
RNN	Recurrent Neural Network.
KNN	K-Nearest Neighbors.
DT	Decision Trees.
RF	Random Forest.
SGD	Stochastic Gradient Descent.
SVM	Support Vector Machine.
AI	Artificial Intelligence.
NLP	Natural Language Processing.
Pos	positive.
Neg	negative.
Neu	neutral.
Comp	compound.
API	Application Programming Interface.
MSE	The mean squared error.
RMSE	The root mean square error.
MAE	The mean average error.

SAE	The simple accuracy error.