ISSN: 2564-6605
doi: 10.28948/ngumuh.535232
Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi, Cilt 8, Sayı 2, (2019), 638-651
Omer Halisdemir University Journal of Engineering Sciences, Volume 8, Issue 2, (2019), 638-651

Araştırma / Research

ANOMALOUS ACTIVITY DETECTION FROM DAILY SOCIAL MEDIA USER MOBILITY DATA

Ahmet Şakir DOKUZ[1] (ORCID: 0000-0002-1775-0954)*

[1]Nigde Omer Halisdemir University, Engineering Faculty, Computer Engineering Department, Nigde, Turkey

Geliş / Received: *04.03.2019*
Kabul / Accepted: *20.05.2019*

ABSTRACT

Anomalous activities are the activities that do not fit into normal and routine behavior of people or objects. Anomalous activity, account, or sharing detection from social networks play an important role for preventing social media users from harmful and annoying contents. However, detecting anomalous activities is challenging due to the difficulty of separating anomalous activities from real ones, limitations of current algorithms and interest measures, the challenge of analyzing social media big data, and hardness of handling spatial and temporal dimensions. In this study, anomalous activities are detected using daily social media user mobility data. In particular, two features are extracted from daily social media user mobility, namely, daily total number of visited locations and daily total distance, and these features are used for detecting anomalous activities. An algorithm, that employs DBSCAN clustering algorithm, is proposed for detecting such activities. The results show that proposed algorithm could learn normal daily activities of social media users and detect anomalous activities.

Keywords: Anomalous activity detection, social networks, spatial social media mining, DBSCAN algorithm

GÜNLÜK SOSYAL MEDYA KULLANICI HAREKETLİLİK VERİLERİNDEN ANORMAL AKTİVİTELERİN TESPİTİ

ÖZ

Anormal aktiviteler, insanlar veya nesnelerin normal ve rutin davranışlarına uymayan aktiviteleri ifade etmektedir. Sosyal ağlardan anormal aktivite, hesap veya paylaşımların tespiti, sosyal medya kullanıcılarını zararlı ve rahatsız edici içeriklerden uzak tutmak için önem taşımaktadır. Ancak anormal aktivitelerin tespiti, anormal aktivitelerin gerçek olanlardan ayrılmasının zor olması, mevcut algoritmalar ve değerlendirme ölçütlerinin yetersiz olması, sosyal medya büyük verisinin analizinin zorlukları ve mekânsal ve zamansal boyutların ele alınmasının zorluklarından dolayı zordur. Bu çalışmada günlük sosyal medya kullanıcı hareketlilik verisi üzerinden anormal aktivitelerin tespiti yapılmıştır. Ayrıntılı olarak, sosyal medya kullanıcı hareketliliklerinden, günlük toplam ziyaret edilen lokasyon sayısı ve günlük toplam uzaklık adında iki özellik çıkarılmış ve bu özellikler anormal aktivitelerin tespitinde kullanılmıştır. Anormal aktivitelerin tespiti için DBSCAN kümeleme algoritmasını kullanan bir algoritma önerilmiştir. Elde edilen sonuçlar önerilen algoritmanın sosyal medya kullanıcılarının normal günlük aktivitelerini öğrenebildiğini ve anormal aktiviteleri tespit edebildiğini göstermiştir.

Anahtar kelimeler: Anormal aktivite tespiti, sosyal ağlar, mekânsal sosyal medya madenciliği, DBSCAN algoritması

*Corresponding author / Sorumlu yazar. Tel.: +90 388 225 2482; e-mail / e-posta: adokuz@ohu.edu.tr

1. INTRODUCTION

Social networks are online social platforms that people communicate with each other, follow status updates of their friends, make new friends, follow thought leaders and eminent people that they give importance [1, 2]. Online social networks provide opportunity to people to provide text, image, and spatial information about their current situation. Several social networks are founded for different purposes, such as, Facebook for connecting past and current friends of people, Twitter for sharing status updates of people with the world, Youtube for sharing video content, Instagram for sharing image content. Social networks are growing in size of the number of users, number of daily activities and also number of connections between users of these social networks. With these characteristics, social networks are one of the main big data sources that have many types of data.

Social media mining is the process of mining social media dataset using data mining methods [3]. The data volume and dimensions of social networks provide vast amount of data to researchers that are working in data mining domain. Three main data mining applications of social networks can be listed as i) semantic analysis of sharings of social media users [4, 5], ii) network analysis between social media users [6, 7], and iii) spatial and spatio-temporal analysis of social media users and sharings [8-10].

Social network users, especially bot accounts, generate anomalous activities for different purposes, such as, trying to appear in search results of a geographical region, simulating real user activities, or pretending to be at a location without actually going there. For these purposes, virtual private networks (VPNs) and other virtual systems are used when sharing from social networks. Social networking sites could not catch these activities, since location information of the users do not seem anomalous individually. Collective analysis of social media user location information could reveal anomalous activities.

Anomalous activity, account, and sharing detection is an important part of social media mining research since detecting anomalous activities, accounts or sharings are essential for providing accurate, and true information to social network users and to protect them from malicious activities [11]. However, detecting these activities is challenging for several reasons. First, many anomalous and malicious activities imitate real user behaviors in social networks and it’s hard to separate anomalous activities from real ones. Second, current algorithms are not suitable for detecting such activities. Third, social media is a big data source and it’s hard to analyze such data. Finally, fourth, social networks have spatial and temporal dimensions that makes detecting anomalous activities harder.

In the literature many studies are conducted to detect anomalous activities from social networks. We can divide these studies into three distinctive categories, such as, contextual, network-oriented, and spatial and spatio-temporal. Contextual anomalous activity detection approaches aim to detect anomalous activities that is based on context and textual information of social media sharings. Network-oriented approaches aim to detect anomalous activities based on their users’ follow relations. Spatial and spatio-temporal approaches aim to detect anomalous activities based on spatial and/or temporal information of social media sharings. Our study falls into the final category, that we aim to detect anomalous activities based on social media users’ daily number of visited locations and the distance between these locations.

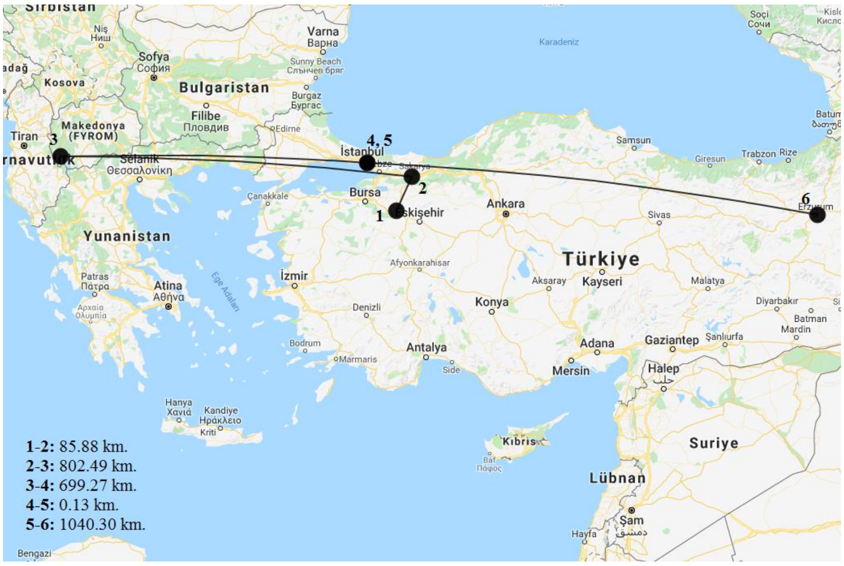

Figure 1. A selected daily anomalous activity of a social media user

Figure 1 presents a sample anomalous daily activity of a social media user. The distances of each sequentially visited locations are also provided bottom-left of the figure. This user visited 6 locations in a day, and one of these locations is to a foreign country, to Albania. Although, locations 1, 2, 4, and 5 could be seen as normal for a daily activity, locations 3 and 6 are distant locations from these four locations. It's not an expected activity for a person to visit Bilecik, Turkey and Sakarya, Turkey and go to Ohri, Albania and return to Istanbul, Turkey and again visit Erzurum, Turkey within a day. Total distance of these 6 locations is 2628.07 km. which is really high for a real activity. This kind of activities should be detected as anomalous activities and the users that share these types of activities should be carefully analyzed.

In this study, anomalous activities of social media users are detected using their daily mobility data. In other words, the number of daily visited locations and the distance of these locations in sequential order are taken into account to determine anomalous activities. An algorithm that uses machine learning clustering algorithm of DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm is proposed that uses these two features, and daily activities of the users are clustered. The daily activities that do not fit into the clusters are determined as anomalous activity. The main contributions of this study are listed as follows:

- Daily mobility data of social media users are used for anomalous activity detection.
- To detect anomalous activities, the number of visited locations and the distance of these locations for each day are used as features.
- A new algorithm that employs DBSCAN algorithm is proposed to cluster daily mobility data of social media users and to detect anomalous activities.
- The algorithm is applied on real social media user dataset and anomalous activities are detected.

The rest of the study is organized as follows. Section 2 presents related work. Section 3 presents the definitions and proposed machine learning based anomalous activity detection algorithm. Section 4 presents the dataset and experimental results. Section 5 presents conclusions.

2. RELATED WORK

The studies related to this study can be divided into three distinctive categories; such as, detecting anomalous activities based on contextual, network-oriented, and spatial and spatio-temporal information. The studies in each of these categories are presented.

Detecting anomalous activities based on contextual information try to detect anomalous activities that is oriented and caused based on context and textual information of social media sharings. Chen et al. [12] proposed two types of anomaly features, namely, domain and social anomaly features and proposed a suspicious URL identification system which is based on Bayesian classification. Takahashi et al. [13] proposed a probabilistic model of the mentioning/replying behavior of a user to detect emergence of a new topic from the anomalies measured through proposed model. Trang et al. [14] and Ruan et al. [15] proposed methods to profile and detect compromised social media accounts. Song et al. [16] proposed a malicious crowdsourcing detection method to discover target objects, such as URLs, search keywords, and posts, by using their manipulation strategy. The authors claim that the proposed method is effective for both spam accounts and real accounts that participate to malicious activity. Aswani et al. [17] proposed a hybrid ABC and k-NN algorithm to identify buzz in Twitter using several metrics, such as, increase in authors, attention level, burstiness level, contribution sparseness, author interaction, author count and average length of discussions are used to model the buzz. Gurajala et al. [18] analyzed 62 million Twitter user data and extracted fake profiles using screen-names, update times, and URLs of fake accounts revealed different characteristics of fake profiles from legitimate users. Gilani et al. [19] and Van der Walt and Eloff [20] proposed methods and algorithms to classify social media users as automated bots and human users using social media account information. These studies focused on discovering anomalous activities based on social features of users and did not take into account spatial information.

Detecting anomalous activities based on network information try to detect anomalous activities based on follow relations of users. Yu et al. [21] proposed GLAD model which is based on hierarchical Bayes model to automatically infer the groups and detect group anomalies simultaneously. Kaur et al. [22] proposed a graph based approach that uses degree and centrality as metrics to spot and rank the anomalous nodes from social networks. Liao et al. [23] presented several visualization techniques for anomaly analysis on large scale social networks that includes community, temporal, correlation, high-dimensional and topology based features. Jeong et al. [24] proposed a classification scheme to detect follow spammers with the assumption that behavior of spam users is different than legitimate users using cascaded social media information of users. Feng et al. [25] proposed *GroupFound* algorithm that focuses on the structure of local graphs other than individual users to detect suspicious accounts. Dutta et al. [26] investigated the retweet behavior of social media users for promoting sharings of other blackmarket service customers by using 64 novel features. Eswaran et al. [27] proposed SpotLight approach which is based on randomized sketching to detect anomalies in streaming graphs. These studies focused on detecting anomalies based on network information of social media and other data sources.

Detecting anomalous activities based on spatial and spatio-temporal information try to detect anomalous activities based on spatial and/or temporal information of social media sharings. Gabrielli et al. [28] proposed new methods to detect urban mobility patterns and anomalies by analyzing social media trajectories and semantically enriching these trajectories. Zheng et al. [29] proposed a method to detected collective anomalies that is composed of three modules, Multiple-Source Latent-Topic, Spatio-Temporal Likelihood Ratio Test, and Candidate Generation by using several spatio-temporal datasets. Chae et al. [30] proposed a trajectory-based visual analytics system for analyzing anomalous human movements during disasters using social media data sources. Pan et al. [31] proposed a method to detect and describe traffic anomalies by using human mobility, such as, GPS trajectories of vehicles, and social media data. Jayarajah et al. [32] focused on detecting anomalies or *events* at city-scale using different urban data sources, such as, traffic cameras, public bus and registered taxi data and Twitter data. Giridhar et al. [33] proposed ClariSense+ system that tries to explain likely causes of traffic anomalies by identifying unusual social network feeds that are correlated with each anomaly in time and in space using vehicular traffic datasets. Huang et al. [34] and Wu et al. [35] proposed systems for predicting urban anomalies as soon as they happen with using spatial and temporal information. Souza et al. [36] proposed a method to detect geographic clusters of dengue infection using spatial and temporal trajectories of social media users. The users are selected by using sentiments of users in their message content. These studies use spatial or spatio-temporal information about different data sources including social media data, however anomaly detection from daily activities of social media users is not studied.

In this study, the problem of anomalous activity detection is investigated. In particular, the use of daily human mobility data is taken into account using two features, namely, daily total number of visited locations and daily total distance of these locations. DBSCAN clustering algorithm is applied on these features to detect anomalous activities of social media users.

3. MATERIAL AND METHOD

In this section, first, the features of detecting anomalous activities from daily mobility data of social media users are provided. Second, machine learning algorithms that are used for clustering user daily mobility data and for detecting anomalous data instances are introduced. Finally, proposed algorithm is presented.

3.1. Daily Mobility Features

In this section two features are extracted from daily social media mobility data to detect anomalous activities based on spatial information of social media users. These features are number of daily visited locations and daily total distance. The calculations of these features are provided as follows.

Number of daily visited locations from social media history of a user is the count of the number of sharings for each day of the user. Daily total distance is calculated as the sum of Euclidean distance of each sequential location. For example, if a user visited locations L1, L2, and L3 in sequential order, then number of daily visited locations is assigned as 3, and daily total distance is the sum of Euclidean distances of (L1-L2) and (L2-L3) pairs. The pseudo-code of calculation of features is presented in Algorithm 1 in Section 3.3. Figure 2 presents graphical representation of proposed features for a sample user.

As can be seen in Figure 2, the user visits 2-15 locations for her/his social media history. Also, daily total distance varies from few to thousands of kilometers. The distribution of daily total distance feature becomes less frequent for higher distance values. Some of the anomalous activities are obviously seen in the figure.

These two features are used because detecting anomalous behavior could happen either with visiting more locations daily, or with visiting distant locations. By using these two features, we aim to detect both abnormal number of location visits and detect abnormal distance of visited locations. We did not use any normalization on these features for the purpose of daily total distance having more effect on clustering performance than number of daily visited locations.

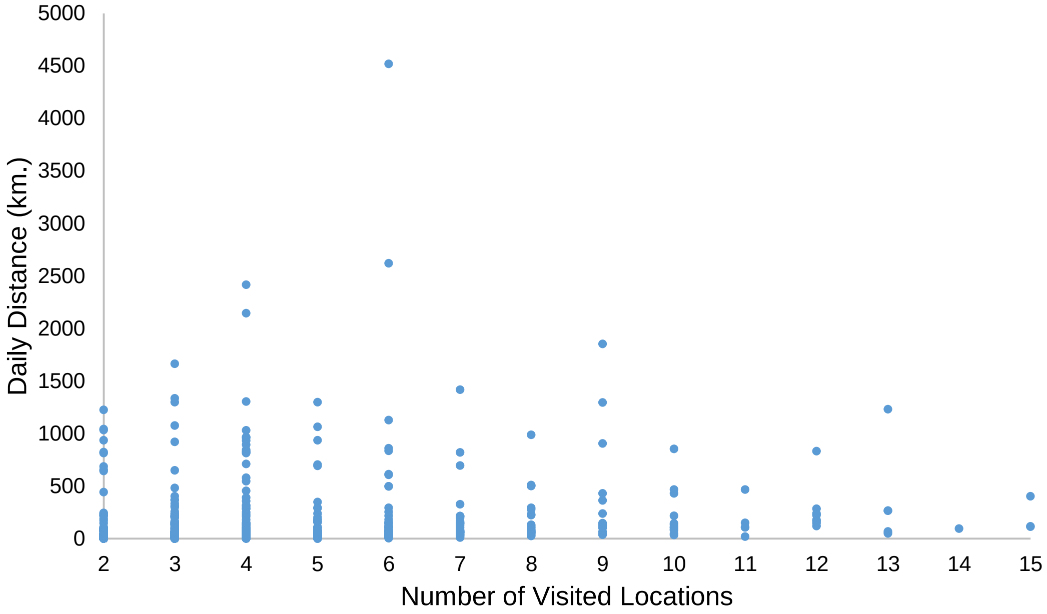

Figure 2. Graphical representation of proposed features for a sample user

3.2. DBSCAN Algorithm

DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm is one of the popular machine learning clustering algorithm that is used for clustering data instances [37, 38]. Alternatively, DBSCAN algorithm can be used for detecting anomalous instances from a set of data points. DBSCAN algorithm is a density based spatial clustering algorithm that could successfully cluster arbitrary shapes of clusters. The main principle of DBSCAN algorithm is to request two user-defined parameters of epsilon (*eps*) and minimum number of points (*min_pts*) and using these parameters to determine clusters. *eps* is used for assigning neighborhood to the data instances, and *min_pts* is used to form data instances into clusters. If data instances are within *eps* distance of an instance, and are at least *min_pts* number of data points are present, then a cluster is formed. DBSCAN algorithm is successfully applied in several application domains, such as, climate [39, 40], bioinformatics [41], renewable energy sources [42], and social media [43].

DBSCAN algorithm has five labels for the data instances, such as, core, border, and noise (anomalous) points [37, 44]. Core points are those that have at least *min_pts* number of points in the *eps* distance. Border points can be defined as points that are not core points, but are the neighbors of core points. Noise points are those that are neither core points nor border points. Figure 3 presents a sample of labelled points from DBSCAN algorithm.

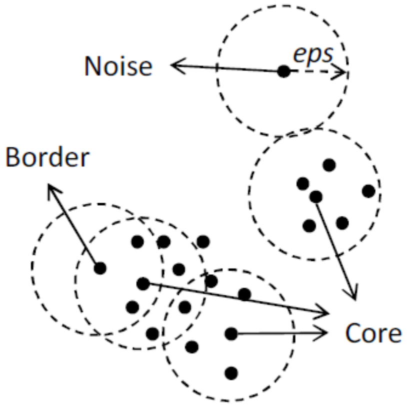

Figure 3. DBSCAN algorithm point types with *eps* = 1 and *min_pts* = 5 [44]

In this study, DBSCAN algorithm is used to determine anomalous activities from proposed two features. Labelled anomalous points are defined as anomalous activity of social media users with respect to number of visited locations and total daily distance. In DBSCAN algorithm, Euclidean Distance is preferred, since we are trying to cluster spatial objects. Different *eps* and *min_pts* values are used to find best parameter set for detecting anomalous daily activities.

3.3. AnomAD Algorithm

In this section, proposed algorithm of AnomAD (**Anom**alous **A**ctivity **D**etection) for detecting anomalous activities is presented. The algorithm is composed of two sub-processes, namely, extracting features for anomalous activity detection, and application of machine learning algorithms for clustering and detecting anomalous activities. The proposed algorithm is presented in Algorithm 1.

Algorithm 1. AnomAD algorithm

Inputs:
D: Social media data of users
eps: Epsilon parameter of DBSCAN algorithm
min_pts: Minimum number of points parameter of DBSCAN algorithm

Output: Detected anomalous activities of social media users

Algorithm:

```
1.  Initialization: numberOfVisitedLocs = null, dailyTotalDist = null, anomalousActivities = null
2.
3.  Prepare features for clustering
4.  for each user u in D
5.    dailyPath = extract-daily-path(u)
6.    for each daily path p in dailyPath
7.      dailyDist = 0
8.      numberOfVisitedLocs ← get-loc-count(p)
9.      for each location pair l in p
10.       dailyDist += euclidean-dist(l)
11.     end for
12.     dailyTotalDist ← dailyDist
13.   end for
14. end for
15.
16. Cluster features to detect anomalous activities
17. for each user u in D
18.   clusters = cluster-dataset(numberOfVisitedLocs, dailyTotalDist, eps, min_pts)
19.   userAnomalousActivities = detect-anomalous-activities(clusters, numberOfVisitedLocs,
    dailyTotalDist)
20.   anomalousActivities ← userAnomalousActivities
21. end for
22. return anomalousActivities
```

Algorithm 1 is organized as two sub-processes. First one is preparing the features for detecting anomalous activities, that is presented at steps 3-14. Second one is applying DBSCAN algorithm and detecting anomalous activities based on features, that is presented at steps 16-21. In first sub-process, at step 5, daily path of users are extracted using extract-daily-path function. In this function, social media history of users is converted into day information and the labels of visited locations. At step 6, each path of user daily path is traversed and the features of number of visited locations and daily total distance are calculated. At step 8, number of daily visited locations are stored in the array of numberOfVisitedLocs. At steps 9-11 distance of each location that user visited in the day is counted using euclidean-dist function and added to dailyDist parameter. Daily total distance of visited locations is stored in the array of dailyTotalDist.

In the second sub-process, each user is traversed and DBSCAN clustering algorithm is applied on the features of each user. After, the anomalous activities are detected using non-clustered instances, in other words anomalous points, of DBSCAN algorithm using detect-anomalous-activities function. Finally, the algorithm returns anomalous activities of each user in the user set.

4. RESULTS AND DISCUSSION

In this section, the performance of proposed AnomAD algorithm is evaluated. First, the dataset and preprocessing steps are introduced, and then the experimental results are presented. The experiments are conducted to reveal runtime performance of proposed algorithm, the impact of DBSCAN algorithm parameters on anomalous activity detection, and the evaluation of results of anomalous behaviors that the algorithm finds. Figure 4 presents experimental setup of this study.

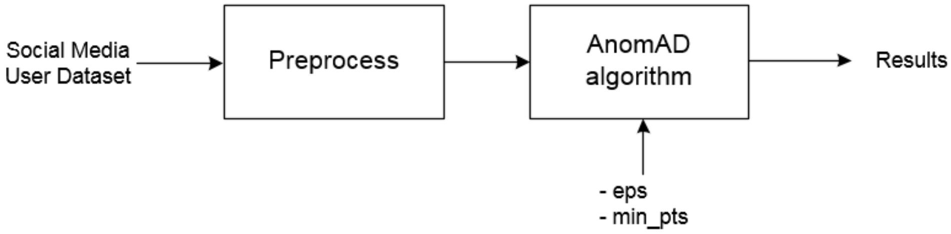

Figure 4. Experimental setup

Experiments are performed to answer following questions:

- What is the effect of the number of users on AnomAD algorithm?
- What is the effect of DBSCAN algorithm parameters if *eps* and *min_pts* on AnomAD algorithm?
- Which anomalous activities are found by AnomAD algorithm?

The experiments are conducted on Intel Core i7 CPU with 3.40 GHz, and 8 GB of RAM.

4.1. Dataset

In this section, first, the social media dataset that is used in this study is presented, and then the preprocessing steps that are applied to the dataset are introduced.

4.1.1. Dataset

In this study, we selected Twitter as a social network and selected dataset have geographical information of social media users. Twitter allows researchers to collect data from their servers and provides several APIs for this purpose. We used REST API and Streaming API [45] for collecting data of Twitter users. Twitter4j [46], an open source Java library, was used for connecting to Twitter APIs and getting user tweets. Java programming language is preferred as a programming language. Twitter Streaming API is used for geographical search on tweets that are posted within the data collection period. Turkey-based geographical search is performed on Twitter Streaming API to get Turkey users as a result. Approximately 2000 users were collected using Streaming API. Next, REST API was used to collect all historical tweets of the users that are collected with Streaming API. Date/time, latitude, and longitude parameters are collected for each tweet of each user.

4.1.2. Preprocessing Steps

In this study, collected users are eliminated using several criteria, i.e. number of tweets, followers count, and followers/friends ratio. We eliminated the users that have lower than 50 tweets. Also, the users, that have followers count less than 10 and followers/friends ratio is below 0.1, are eliminated from the dataset. These criteria are used for many spam user detection literature and detailed information can be found in [47] and [48].

After the preprocessing step, random 1000 users were selected for the experiments from remaining user set.

4.2. Experimental Results

In this section, the experiments of detecting anomalous activities from user daily mobility data are presented. Three experiments are performed to reveal performance of proposed algorithm, the effect of number of users, effect of DBSCAN algorithm parameters, and the evaluation of the results.

4.2.1. Effect of Number of Users

In this experiment, runtime performance of proposed AnomAD algorithm is evaluated whether it’s scalable with respect to the increase of the dataset. *eps* and *min_pts* parameters are set at 3, and 6, respectively. Number of users is increased from 200 to 1000 by 200 and runtime of the algorithm for each value is extracted. Effect of number of users is presented in Figure 5.

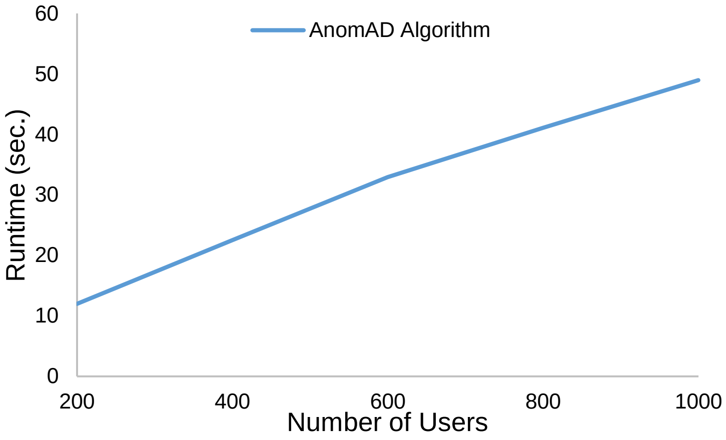

Figure 5. Effect of number of users on AnomAD algorithm

As can be seen in Figure 5, proposed AnomAD algorithm could detect anomalous activities from social media user dataset within reasonable time. For the whole dataset of 1000 users, the algorithm provided results approximately 50 seconds. Also, AnomAD algorithm proved to be linearly scalable to the increase of the dataset size.

4.2.2. Effect of DBSCAN Algorithm Parameters

In this experiment, the effect of DBSCAN algorithm parameters are evaluated, which are *eps* and *min_pts*. We evaluated the effect of these parameters as detected anomalous activity count. Number of users are set at 1000. *eps* parameter is set at 3 for the analysis of *min_pts* parameter, and *min_pts* parameter is set at 6 for the analysis of *eps* parameter. It is expected to decrease number of anomalous points when *eps* parameter increases and *min_pts* parameter decreases, and to increase number of anomalous points vice versa. It is essential to set these parameters carefully because low and high parameter values cause wrongly detected anomalous activities. Figure 6 a-d presents effect of *eps* and *min_pts* parameters with respect to total anomalous activity counts.

As can be seen in Figure 6 a-b, the increase of *eps* parameter decreases anomalous activity counts. The main reason for this is, when *eps* parameter increases, the radius to be inside same cluster increases and actually distant points are put into the same cluster. Contrarily as can be seen in Figure 6 c-d, the increase of *min_pts* parameter increases anomalous activity counts. The main reason for this behavior is, when *min_pts* parameter increases, the rule to be formed and treated as a cluster gets harder and more points are classified as anomaly points.

Another outcome of the analysis results is mean anomalous activity counts. When Figure 6 b and d are analyzed, at least 10 anomalous activities are present for each social media user of the dataset. This value is reasonable since social media users tend to share their locations when they move off their daily expected patterns.

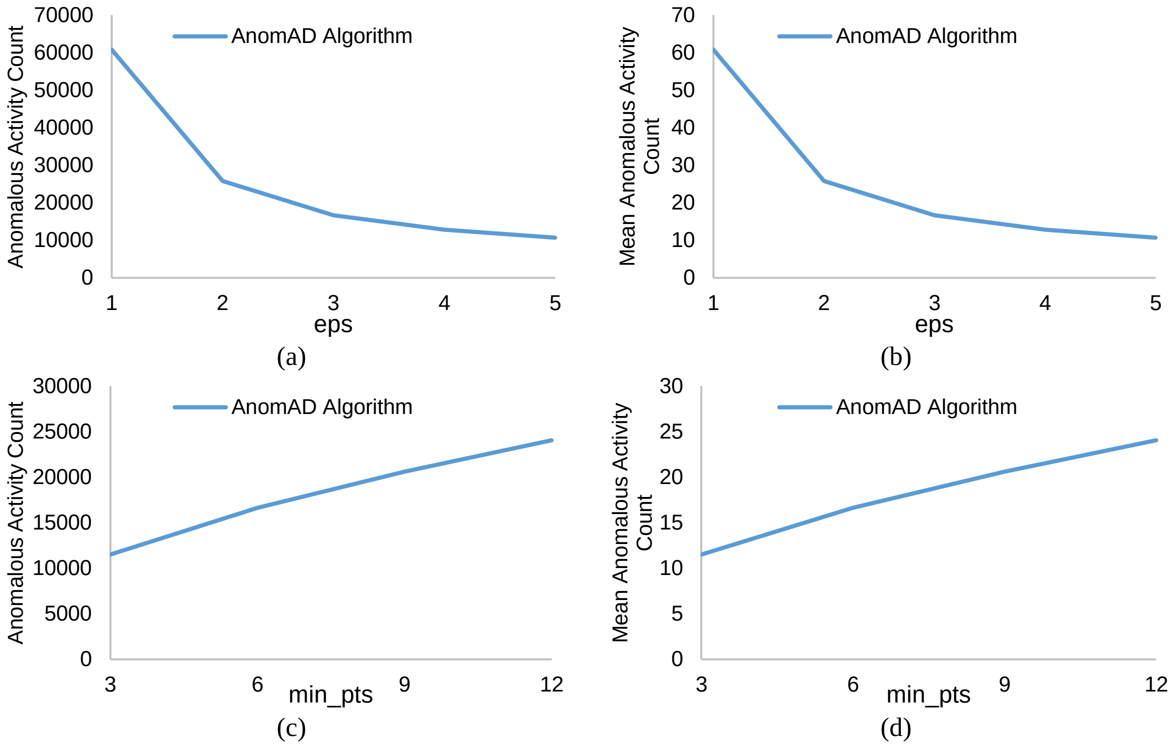

Figure 6. Effect of DBSCAN algorithm parameters on AnomAD algorithm

4.2.3. Evaluation of Detected Anomalous Activities

In this experiment, detected anomalous activities that AnomAD algorithm finds are evaluated. Number of users, *eps* and *min_pts* are set at 1000, 3, and 6, respectively. First, top 10 users that have highest number of anomalous activities are presented, then, first user (User 1 in Table 1) is statistically and graphically investigated. Then, least number of anomalous activities are evaluated. Finally, a histogram of number of anomalous activities is provided.

Table 1. Top 10 users that have highest number of anomalous activities

Order	Number of Anomalous Activities	Tweet Count	Anomalous Activity Ratio
User 1	190	3194	**0.059**
User 2	130	1969	**0.066**
User 3	115	3187	0.036
User 4	98	3195	0.031
User 5	85	3224	0.026
User 6	82	1398	**0.059**
User 7	78	3120	0.025
User 8	77	3164	0.024
User 9	76	1000	**0.076**
User 10	74	3195	0.023

As can be seen in Table 1, among 1000 social media users, these 10 users have the highest number of anomalous activities. First user has 190 anomalous activities among 3194 tweets, which is a high proportion of anomalous activities. Four of these 10 users, Users 1, 2, 6, and 9, have higher anomalous activity ratio with respect to other users, which means that these users are producing more anomalous activities. An interesting result is obtained at User 9. This user has relatively low number of anomalous activities with respect to other users in Table 1, however, this user has highest anomalous activity ratio, that is 0.076. This means that User 9 shares low number of tweets, but the probability of anomalous activity of these tweets are more expected. The statistical properties of User 1 are presented in Table 2.

Table 2. Statistical properties of anomalous activities of User 1 at Table 1

Number of Visited Locations	Average Daily Distance of Anomalous Activities (km.)	Maximum Daily Distance of Anomalous Activities (km.)	Average Daily Distance of Normal Activities (km.)	Maximum Daily Distance of Normal Activities (km.)
2	438.75	1225.99	16.03	106.17
3	364.41	1665.12	28.47	105.20
4	635.47	2418.28	43.44	151.45
5	461.31	1299.30	40.62	109.90
6	803.17	**4519.23**	52.40	151.99
7	384.92	1417.82	64.50	152.19
8	329.63	988.87	62.13	108.15
9	479.88	1853.07	53.54	65.34
10	249.97	855.67	108.39	109.42
11	212.18	467.24	108.10	108.40
12	273.97	833.56	-	-
13	376.91	1233.20	-	-
14	96.15	96.15	-	-
15	211.77	403.25	-	-

As can be seen in Table 2, number of visited locations per day are 2-15. Average and maximum distances of anomalous and normal activities are presented for each number of visited location. It’s evident that average and maximum daily distances of anomalous activities are quite higher than normal activities. For example, for 2 visited locations, first row in the table, average daily distance of normal activities are 16.03 km, however average daily distance of anomalous activities are 438.75 km. Also, maximum daily distance of 2 visited locations is 1225.99 km which might not be a real activity. For 6 visited locations, the maximum daily distance of anomalous activity is observed as 4519.23 km, that is also considered as not a real activity. Another observation from Table 2 is that as the number of visited locations increase, the gap between average and maximum daily distance of normal activities are getting smaller. Contrarily, average and maximum daily distance of anomalous activities have no linear correlation. There is no cluster for number of visited locations 12, 13, 14, and 15, and thus normal activity distance values are not present. As a result, when detected anomalous activities are evaluated for a user, the algorithm can be said successful for detecting anomalous activities.

Table 3. Selected anomalous activities of User 1

Order	Daily Path	Number of Visited Locations	Total Daily Distance
1	Belgrade, Serbia – Ankara, Turkey – Ankara, Turkey – Belgrade, Serbia – Ankara, Turkey – Belgrade, Serbia	6	4519.23
2	Bilecik, Turkey – Sakarya, Turkey – Ohri, Albania – Istanbul, Turkey – Istanbul, Turkey – Erzurum, Turkey	6	2628.07
3	Istanbul, Turkey – Istanbul, Turkey – Adana, Turkey – Adana, Turkey – Mersin, Turkey – Mersin, Turkey – Mersin, Turkey – Adana, Turkey – Adana, Turkey – Mersin, Turkey	10	855.67
4	Istanbul, Turkey – Kocaeli, Turkey – Mugla, Turkey – Kocaeli, Turkey – Kocaeli, Turkey – Istanbul, Turkey – Istanbul, Turkey – Mugla, Turkey – Kocaeli, Turkey	9	1853.07
5	Istanbul, Turkey – Kocaeli, Turkey – Vanadzor, Armenia – Bolu, Turkey	4	2418.28
6	Ankara, Turkey – Duzce, Turkey – Mugla, Turkey – Sakarya, Turkey – Kocaeli, Turkey – Kocaeli, Turkey – Kocaeli, Turkey – Kocaeli, Turkey – Kocaeli, Turkey – Kocaeli, Turkey – Kocaeli, Turkey – Kocaeli, Turkey – Kocaeli, Turkey	13	1233.20
7	Kocaeli, Turkey – Kocaeli, Turkey – Kocaeli, Turkey – Mugla, Turkey – Kocaeli, Turkey – Istanbul, Turkey – Kocaeli, Turkey	8	988.87
8	Ohri, Albania – Istanbul, Turkey – Kayseri, Turkey	3	1297.54
9	Belgrade, Serbia – Istanbul, Turkey – Ohri, Albania – Istanbul, Turkey	4	2145.52
10	Istanbul, Turkey – Ankara, Turkey – Erzurum, Turkey – Erzurum, Turkey	4	1029.87

Table 3 presents selected anomalous activities for the sample user, User 1. As can be seen in daily paths of Table 3, the user has a characteristic of visiting same locations, i.e. cities, in sequential order. For example, at the first anomalous activity, the user visited Belgrade, Serbia and Ankara, Turkey in a sequential order for 3 times for each location. This activity could not be a real activity, since a user might not visit a different country more than twice within a day. Similar patterns are observed at anomalous activities of 3, 4, 7, and 9. Some of the activities that are presented at Table 3 could be real activities, such as activities 5, 8, or 10. These activities could be real activities, however these activities are far beyond the normal activity of the user. For this reason, it is also important to detect such activities for many application areas, such as tourism destinations recommendation, hotel and restaurant advertisements, or providing information about the new environment.

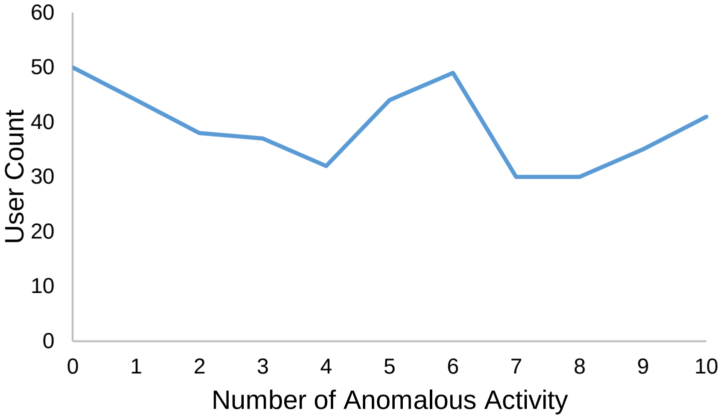

Figure 7. Number of users that have few number of anomalous activities

Figure 7 presents number of users that have anomalous activities smaller than or equal to 10. As can be seen in Figure 7, 50 users have no anomalous activities, which means that these users are sending tweets inside a region or city and do not have abnormal activity. The number of users for each number of anomalous activity is within the range of 30-50.

A histogram of number of users with respect to number of anomalous activities is given in Figure 8. As can be seen in Figure 8, the number of users that have smaller than or equal to 10 anomalous activities are 430 which means that nearly half of total 1000 social media user group have reasonable number of anomalous activities. Another observation is that number of users that have more anomalous activities are getting smaller as the number of anomalous activities increases. There are only 75 users that have more than 41 anomalous activities.

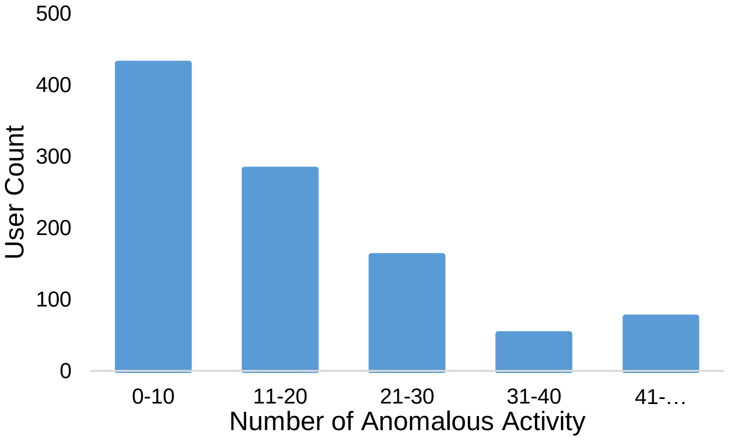

Figure 8. Histogram of user count with respect to number of anomalous activities of the users

Based on the results that AnomAD algorithm finds, it can be said that the algorithm could successfully detect anomalous activities from social media user daily mobility data using number of visited locations and daily distances of these locations.

5. CONCLUSION

Detecting anomalous activities, accounts or sharings is an important part of social media mining research since detecting such activities, accounts or sharings are essential for providing accurate and true information to social network users and to protect them from malicious activities. However, detecting these activities is challenging due to the difficulty of separating anomalous activities from real ones, limitations of current algorithms and interest measures, the challenge of analyzing big data, and hardness of handling spatial and temporal dimensions. In this study, anomalous activities are detected using daily social media user mobility data. In particular, two features are extracted from daily mobility data of social media users, namely, daily total visited locations, and daily total distance, and clustering algorithm of DBSCAN is applied on these features to detect anomalous activities of social media users. AnomAD algorithm, that employs DBSCAN clustering algorithm, is proposed for detecting such activities. The results show that proposed algorithm could learn normal daily activities and detect anomalous activities.

For future studies, AnomAD algorithm could be extended to use other clustering algorithms. This study detects anomalous daily activities rather than analyzing each activity alone, thus a novel method could be developed that takes into account each daily activity apart from daily activities of social media users. Also, daily activity features could be increased to enhance anomalous activity detection performance. Finally, the performance of the algorithm could be tested on other data sources, such as, GPS, or urban datasets.

REFERENCES

[1] JIN, L., CHEN, Y., WANG, T., HUI, P.,VASILAKOS, A. V., "Understanding user behavior in online social networks: a survey", IEEE Communications Magazine, 51*:* 144-150, 2013.

[2] YU, R., QIU, H., WEN, Z., LIN, C.,LIU, Y., "A Survey on Social Media Anomaly Detection", SIGKDD Explor. Newsl., 18*:* 1-14, 2016.

[3] ZAFARANI, R., ABBASI, M. A., LIU, H., " Social Media Mining: An Introduction", Cambridge University Press, p. 332, 2014.

[4] PENG, H., BAO, M., LI, J., BHUIYAN, M. Z. A., LIU, Y., HE, Y., YANG, E., "Incremental Term Representation Learning for Social Network Analysis", Future Generation Computer Systems, 86*:* 1503-1512, 2018.

[5] ECEMIŞ, A., DOKUZ, A. Ş., ÇELIK, M., "Sentiment Analysis of Posts of Social Media Users in Their Socially Important Locations", *2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2015, pp. 1-6.*

[6] RONG, H., MA, T., TANG, M., CAO, J., "A novel subgraph K$^+$-isomorphism method in social network based on graph similarity detection", Soft Computing, 22(8)*:* 2583-2601, 2018.

[7] LEI, K., LIU, Y., ZHONG, S., LIU, Y., XU, K., SHEN, Y., YANG, M., "Understanding User Behavior in Sina Weibo Online Social Network: A Community Approach", IEEE Access, 6*:* 13302-13316, 2018.

[8] SCELLATO, S., NOULAS, A., LAMBIOTTE, R., MASCOLO, C., "Socio-spatial properties of online location-based social networks", *Fifth International AAAI Conference on Weblogs and Social Media, 2011, pp. 329-336.*

[9] DOKUZ, A. S.,CELIK, M., "Discovering socially important locations of social media users", Expert Systems with Applications, 86*:* 113-124, 2017.

[10] CELIK, M.,DOKUZ, A. S., "Discovering socio-spatio-temporal important locations of social media users", Journal of Computational Science, 22*:* 85-98, 2017.

[11] SAVAGE, D., ZHANG, X., YU, X., CHOU, P.,WANG, Q., "Anomaly detection in online social networks", Social Networks, 39*:* 62-70, 2014.

[12] CHEN, C.-M., GUAN, D. J.,SU, Q.-K., "Feature set identification for detecting suspicious URLs using Bayesian classification in social networks", Information Sciences, 289*:* 133-147, 2014.

[13] TAKAHASHI, T., TOMIOKA, R.,YAMANISHI, K., "Discovering Emerging Topics in Social Streams via Link-Anomaly Detection", IEEE Transactions on Knowledge and Data Engineering, 26*:* 120-130, 2014.

[14] TRÅNG, D., JOHANSSON, F.,ROSELL, M., Evaluating Algorithms for Detection of Compromised Social Media User Accounts. *Proc. 2015 Second European Network Intelligence Conference, 2015,pp. 75-82*

[15] RUAN, X., WU, Z., WANG, H.,JAJODIA, S., "Profiling Online Social Behaviors for Compromised Account Detection", IEEE Transactions on Information Forensics and Security, 11*:* 176-187, 2016.

[16] SONG, J., LEE, S.,KIM, J., "CrowdTarget: Target-based Detection of Crowdturfing in Online Social Networks", Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 793-804, Denver, Colorado, USA, 2015.

[17] ASWANI, R., GHRERA, S. P., KAR, A. K.,CHANDRA, S., "Identifying buzz in social media: a hybrid approach using artificial bee colony and k-nearest neighbors for outlier detection", Social Network Analysis and Mining, 7*:* 38, 2017.

[18] GURAJALA, S., WHITE, J. S., HUDSON, B., VOTER, B. R.,MATTHEWS, J. N., "Profile characteristics of fake Twitter accounts", Big Data & Society, 3*:* 2053951716674236, 2016.

[19] GILANI, Z., KOCHMAR, E.,CROWCROFT, J., "Classification of Twitter Accounts into Automated Agents and Human Users", Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, 489-496, Sydney, Australia, 2017.

[20] WALT, E. V. D.,ELOFF, J., "Using Machine Learning to Detect Fake Identities: Bots vs Humans", IEEE Access, 6*:* 6540-6549, 2018.

[21] YU, R., HE, X.,LIU, Y., "GLAD: Group Anomaly Detection in Social Media Analysis", ACM Trans. Knowl. Discov. Data, 10*:* 1-22, 2015.

[22] KAUR, R., KAUR, M.,SINGH, S., "A Novel Graph Centrality Based Approach to Analyze Anomalous Nodes with Negative Behavior", Procedia Computer Science, 78*:* 556-562, 2016.

[23] LIAO, Q., SHI, L.,WANG, C., "Visual analysis of large-scale network anomalies", IBM Journal of Research and Development, 57*:* 13:11-13:12, 2013.

[24] JEONG, S., NOH, G., OH, H.,KIM, C.-K., "Follow spam detection based on cascaded social information", Information Sciences, 369*:* 481-499, 2016.

[25] FENG, B., LI, Q., PAN, X., ZHANG, J.,GUO, D., "GroupFound: An effective approach to detect suspicious accounts in online social networks", International Journal of Distributed Sensor Networks, 13*:* 1550147717722499, 2017.

[26] DUTTA, H. S., CHETAN, A., JOSHI, B.,CHAKRABORTY, T., Retweet Us, We will Retweet You: Spotting Collusive Retweeters Involved in Blackmarket Services. *Proc. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018,pp. 242-249*

[27] ESWARAN, D., FALOUTSOS, C., GUHA, S.,MISHRA, N., "SpotLight: Detecting Anomalies in Streaming Graphs", Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &#38; Data Mining, 1378-1386, London, United Kingdom, 2018.

[28] GABRIELLI, L., RINZIVILLO, S., RONZANO, F.,VILLATORO, D., From Tweets to Semantic Trajectories: Mining Anomalous Urban Mobility Patterns. *Proc. International Workshop on Citizen in Sensor Networks, 2013,pp. 26-35*

[29] ZHENG, Y., ZHANG, H.,YU, Y., "Detecting collective anomalies from multiple spatio-temporal datasets across different domains", Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, 1-10, Seattle, Washington, 2015.

[30] CHAE, J., CUI, Y., JANG, Y., WANG, G., MALIK, A.,EBERT, D. S., Trajectory-based visual analytics for anomalous human movement analysis using social media. *Proc. EuroVis Workshop on Visual Analytics (EuroVA), 2015,pp. 1-5*

[31] PAN, B., ZHENG, Y., WILKIE, D.,SHAHABI, C., "Crowd sensing of traffic anomalies based on human mobility and social media", Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 344-353, Orlando, Florida, 2013.

[32] JAYARAJAH, K., SUBBARAJU, V., WEERAKOON, D., MISRA, A., TAM, L. T.,ATHAIDE, N., Discovering anomalous events from urban informatics data. *Proc. Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR VIII, 2017 10190,pp. 101900F*

[33] GIRIDHAR, P., AMIN, M. T., ABDELZAHER, T., WANG, D., KAPLAN, L., GEORGE, J.,GANTI, R., "ClariSense+: An enhanced traffic anomaly explanation service using social network feeds", Pervasive and Mobile Computing, 33*:* 140-155, 2016.

[34] HUANG, C., WU, X.,WANG, D., "Crowdsourcing-based Urban Anomaly Prediction System for Smart Cities", Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, 1969-1972, Indianapolis, Indiana, USA, 2016.

[35] WU, X., DONG, Y., HUANG, C., XU, J., WANG, D.,CHAWLA, N. V., UAPD: Predicting Urban Anomalies from Spatial-Temporal Data. *Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2017,pp. 622-638*

[36] SOUZA, R. C. S. N. P., ASSUNÇÃO, R. M., OLIVEIRA, D. M., NEILL, D. B.,MEIRA, W., "Where did I get dengue? Detecting spatial clusters of infection risk with social network data", Spatial and Spatio-temporal Epidemiology2018.

[37] ESTER, M., KRIEGEL, H.-P., #246, SANDER, R.,XU, X., "A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 226-231, Portland, Oregon, 1996.

[38] OZKOK, F. O.,CELIK, M., "A New Approach to Determine Eps Parameter of DBSCAN Algorithm", 2017, 5*:* 5, 2017.

[39] ÇELIK, M., DADAŞER-ÇELIK, F.,DOKUZ, A. Ş., Anomaly detection in temperature data using DBSCAN algorithm. *Proc. 2011 International Symposium on Innovations in Intelligent Systems and Applications, 2011,pp. 91-95*

[40] DEY, R.,CHAKRABORTY, S., Convex-hull &amp; DBSCAN clustering to predict future weather. *Proc. 2015 International Conference and Workshop on Computing and Communication (IEMCON), 2015,pp. 1-8*

[41] EDLA, D. R.,JANA, P. K., "A Prototype-Based Modified DBSCAN for Gene Clustering", Procedia Technology, 6*:* 485-492, 2012.

[42] DOKUZ, A. S., DEMOLLI, H., GOKCEK, M.,ECEMIS, A., "Year-Ahead Wind Speed Forecasting using a Clustering-Statistical Hybrid Method", CIEA' 2018 International Conference on Innovative Engineering Applications, 971-975, Sivas, Turkey, 2018.

[43] LIU, J., HAO, F.,WANG, Y., Discovering Influential Areas According to Check-In Records and User Influence in Social Networks. *Proc. 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2018,pp. 1151-1155*

[44] REZAEI, M.: 'Clustering validation', University of Eastern Finland, 2016

[45] https://developer.twitter.com/en/docs, (erişim tarihi 04.03.2019)

[46] http://twitter4j.org/en/index.html, (erişim tarihi 04.03.2019)

[47] BENEVENUTO, F., MAGNO, G., RODRIGUES, T.,ALMEIDA, V., "Detecting Spammers on Twitter", CEAS 2010 - Seventh annual Collaboration, Electronic messaging, AntiAbuse and Spam Conference, 1-9, Redmond, Washington, USA, 2010.

[48] ZHENG, X., ZENG, Z., CHEN, Z., YU, Y.,RONG, C., "Detecting spammers on social networks", Neurocomputing, 159*:* 27-34, 2015.