An Exploratory Analysis of Leaked Facebook Data: A Case of Turkish Users

Önder Çoban

Adıyaman University, Computer Engineering Department, Adıyaman, Turkey
e-mail: onder.cbn@gmail.com
Research Paper Received: 09.11.2021 Revised: 08.12.2021 Accepted: 08.12.2021

Abstract—In this study, an analysis is performed on recently leaked data of Turkish Facebook users to inspect their sharing behavior along with how accurately an adversary can perform attacks to learn the gender and username of a user. Experimental results show that the majority of users do not disclose their sensitive data except for phone numbers. Users mostly live in big cities, but privacy-aware users mostly live in the eastern and southeastern parts of the country. It is also possible to infer gender with very high accuracy up to 0.95 just using the first name and username of a user.

Keywords—Online social networks, facebook, privacy, sharing analysis.

1. Introduction

In today's world, Online Social Networks (OSNs) are very popular communication tools that enable users to build friendships, share anything about what happens in their lives, follow news and politicians they supported, and play games among many others. Popular OSNs like Facebook, Twitter, and Instagram serve free of charge and connect hundreds of millions of users around the world. As such, OSNs provide a platform for their users for fun and diversion and they also help users to promote their product over the internet, staying connected, and spreading information in a faster way around the world [1]. The use of OSNs has reached such an enormous scale that it rivals the many popular search engines in terms of usage [2]: the number of people around the world using OSNs is expected to grow from 3.6 billion in 2020 to 4.4 billion in 2025 [3]. The world map of OSNs in Figure 1 verifies this truth by showing how deeply OSNs have penetrated people's lives and transformed the ways they communicate.

As a result of this popularity, OSNs vary a lot and there exist a large number of OSNs of different categories like online sharing, networking, video uploading, and so on [4]. Mainly, OSNs are grouped into four categories such as connection, professional purpose, multimedia, and academic [5]. For instance, among the most prominent connection OSNs, Facebook is often preferred by users who want to connect with family members and friends, while Twitter serves as a micro-blogging tool for users who want to read or write the latest news [6]. On the other hand, users in a specific group or occupation use

professional purpose Linked-In to interact and give details about their professional evolution [5], [6].

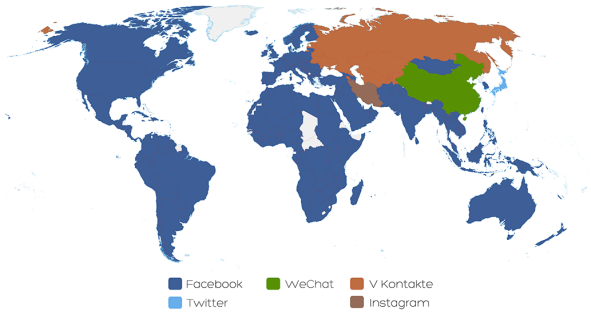


Fig. 1: World map of OSNs as of January 2021 [7]

It is well known that OSNs ask their users for their information and many of those users often provide their information consciously or unconsciously to increase their digital existence in OSNs [8]. This situation paves the way for OSNs to store a great amount of information about users such that their confidentiality cannot be guaranteed [9]. This brings various privacy and security problems to come into view for OSN users since OSN service providers collect the private and sensitive data of their users that can be misused by data collectors, third parties, or unauthorized users [10], [11]. The problem of privacy and security is such serious that users may still be at risk even though they do not disclose their sensitive information [12]. Primary causes of the privacy and security implications include cyberstalking, phishing, cross-site scripting, fake-profiling, and so on [11], [13], [14].

As such, OSNs allow users to control and customize which of their personal information is public to other users or applications [11], [15]. However, third parties can have direct or indirect access to the OSN data in different ways (e.g., crawling and scraping) and they can use or sell it to others with unknown malicious intensions [16]. Besides, even in some cases, OSN service providers can also violate the privacy of their users. For instance, in the Cambridge Analytica scandal, profile information of more than 87 million Facebook users may have been acquired and used to build a software program to predict and influence voters [17]. It is possible to extend similar cases that include revealing private data of 120 million Facebook users [18], putting up Twitter users' trillions of tweets for sale [19], and so on. In a much recent case, basic profile information of 533 million Facebook users is leaked [20] in a public database. This superset of the data contains information of users from 106 countries including Turkey.

In this study, an analysis is performed on a subset of this superset of the leaked data that contains the basic information of approximately 20M Turkish Facebook users. The contributions of this study can be summarized as follows:

- the data is analyzed to discover sharing behavior of a very large volume of users along with how accurately it is possible to perform inference attacks when an adversary has access to a very small part (i.e., just a few of basic profile attributes of users) of OSN data,
- it is tried to infer gender and username attributes of users by using two simple inference mechanisms from which the gender inference mechanism achieves very high accuracy up to 0.95 just using first name and username of a user,
- geographical distribution of users is performed to inspect different cases including where privacy unaware users often live in, and
- findings of this study have great importance in showing the general picture of the current situation of Turkish Facebook users.

The rest of this article is structured as follows: Section 2 reviews the literature with a focus on studies focusing on Turkish OSN users. Section 3 and Section 4 introduce the used materials and methods respectively. Section 5 outlines the experimental results. Section 6 provides a discussion of the results, and finally, Section 7 gives conclusion of this study.

2. Literature Review

With the great increase in the use of OSNs, research focusing on OSN data has attracted the attention of researchers from different disciplines including computer science. In this section, a review of the literature with a focus on studies aiming at crawling, sharing analysis, attribute inference, and privacy analysis is performed especially considering Turkish OSN users.

The published studies that we are aware of are as follows: matching users across multiple OSNs is performed in [10] just by using usernames. The authors state that they achieved an F1 score of 0.92 without feature selection and extension. A privacy risk evaluation of Turkish Facebook users is conducted in [11] by using two state-of-the-art techniques. The results of this study showed that male users and users in the age range 21-40 are at greater risk. Gender inference for Turkish Facebook users is performed based on profile information, social connections, and wall contents in [13]. In this study, the highest accuracy is obtained as an accuracy of 0.98 by using the profile information of users. In [21], Facebook data of 20K users is automatically crawled by visiting seed user's friends in a Breadth-first search (BFS) order. Statistical analysis of the crawled data showed that users generally do not tend to disclose their sensitive attributes either consciously or unconsciously. However, sharing rates of some attributes such as birth date, family members, relationships, place, and work are too high to be underestimated. A new sensitivity computing method is proposed in [22] to use it in privacy risk scoring in OSNs. In this study, experiments performed on both synthetic and real-world datasets imply that there is a strong relation between term weighting process of classical text categorization and privacy risk scoring. In [23], it is tried to detect kinship between two Turkish Facebook users based on a lexicon-based approach that completely relies on wall contents. The authors obtained an F1 score of 0.41 even though their content-based approach has some challenges. In [14], real data of 5,389 Linked-In users from Turkey is crawled to analyze the privacy attitude of users. The results of the analysis showed that location, working experience, education, and area of interest information are the most disclosed ones by users. Similarly, an investigation of the privacy attitudes of Turkish information professionals is conducted in [24], in which the authors concluded that users are often aware of privacy, and most of them change the default settings to protect their personal information. On the other hand, privacy-related consequences of Turkish citizen database leak are inspected in [25]. The results of this study showed that with automated processing of the data, an adversary can uniquely identify the mother's maiden name of individuals and landline numbers for a significant portion of people. In [26], an analysis on the data of 200 popular Turkish companies is performed to detect relationships and similarities among that companies.

According to our extensive review of the literature, the studies summarized above are only the ones dedicated to using real-world data of Turkish users. Notice that the majority of the

existing studies rely on synthetic or surveyed data since accessing real-world data is very hard. According to [21], the primary reasons behind the scarcity of real-world OSN data are privacy concerns, complexity and volume of OSN data, and the value of OSN data. Nevertheless, there exist many other studies on real-world OSN data of Turkish users. However, these studies handle the data with a different point of view and they have a purpose other than the purposes of both this study and studies summarized above. Detection of sentiment [27], [28], [29], abusive[30] and offensive [31] language, cyberbullying [32], [33], event [34], stance [35], and political view [36] to list a few examples that often rely on the textual content of OSN users. Likewise to the [13], content-based approaches are also used for attribute inference in OSNs. In the literature, the previous studies considering Turkish users on this topic also mainly focus on the detection of the gender attribute. In [37], the content of Turkish users' tweets are used to predict their gender. The authors achieved an accuracy of 0.87 for Turkish on a dataset that includes 3.6K Turkish users' tweets. In [38] and [39], the gender of Turkish Twitter users is detected using tweet contents with an accuracy of 0.80 and 0.72, respectively. Two other studies [40], [41] also performed gender detection based on the content of Facebook comments, and the authors obtained an accuracy of 0.90, and 0.74 respectively.

From the literature, it can be understood that accessing OSN data is very hard and the majority of existing studies use synthetic or surveyed data. In the context of Turkish users, many studies are focusing on real-world OSN data, but the majority of them often use a content-based approach to perform text mining tasks such as sentiment analysis. On the other hand, studies focusing on Turkish users' privacy and sharing analysis along with attribute inference are still very limited. Besides, the volume of the data used in the existing studies is generally not quite large.

In this study, exploratory analysis is conducted on recently leaked real-world Facebook data [20] of Turkish users. Even though this data does not contain an underlying graph structure and only contains a small part of users' data, it has a very high large volume (approx. 20M) of users compared to the existing studies. The findings of this study, therefore, have great importance to show that the leaked data can provide even more information than what is visible to the naked eye.

3. Material

3.1. Dataset

The dataset underlying this study includes Facebook data [20] of approximately 20M users from Turkey. This is a subset of the superset of the leaked data that contains data of 533 million Facebook users from 106 countries. It just contains users' eight profile attributes like phone number, first name, surname, email, date of birth, gender, hometown, and lived-in place along with the Facebook ID. According to Facebook, data was leaked in a breach sometime before August 2019 and was recently made available on April 2021 in a public database [42]. Facebook also claimed that it is found and fixed the issue in August 2019 and the same route can no longer be used to scrape that data.

The leaked data set has been posted on the hacking forum for free, making it available to anyone with rudimentary data skills. In this study, the data of Turkish users is obtained to perform an analysis only for academic purposes.

After completing the analysis, the data has been destroyed to prevent possible cases of privacy concerns.

3.2. Lexicons

In this study, a lexicon of Turkish person names is used in simple inference attacks devoted to predicting the gender of users (see Section 5.2.1). This lexicon [43] includes 9,704 Turkish person names of which 989 of them are unisex, while 6,085 and 2,630 of them are male and female names respectively.

Besides, a lexicon of cities and districts of Turkey is used to detect whether a user is from Turkey or not, and if so, in which city (or district) he/she lives and which one is his/her hometown. These steps are employed to extract geographical distribution (see Section 5.1.4) of users across cities of Turkey. This lexicon [44] includes names of 81 cities of Turkey along with district names within each city.

3.3. Geospatial Data

A geospatial data is a database of geographic data, such as countries, administrative divisions, cities, and related information. In this study, geospatial data of Turkey is used to visualize the geographical distribution of users w.r.t. their attributes like hometown and places they lived-in along with their privacy-awareness. Geospatial data of the Turkey (https://gadm.org/maps/TUR.html) is obtained from GADM (The Database of Global Administrative Areas) database (https://gadm.org/) that provides maps and spatial data for all countries and their sub-divisions.

4. Methods

In this study, several methods are used to perform a complete analysis of the leaked data. These methods are chosen simply regarding the attributes and their properties. Since the attribute values (except for the Facebook ID) are comprised of strings, the analysis is mainly based on a few simple string processing methods such as split, replace, regular expresisons, and so on. For instance, a regular expression (https://emailregex.com/) considering the RFC5322 (https://datatracker.ietf.org/doc/html/rfc5322) standard internet message format is used to check that whether an email address valid or invalid. Besides, it is needed to use aggregation functions along with some additional lexicons to group users based on their attributes like hometown, gender, and so on. The following sub-headings give details of methods specifically used to perform other major steps of the analysis.

4.1. Getting candidate substrings

This method is used to extract possible candidate names from a given username when inferring a user's gender based on his/her first name is not applicable (see Section 5.2.1). It basically iterates over the given username within a finite loop, in which at each iteration it gets a substring with a length between s and e that stand for starting and ending positions of the substring respectively. This process is applied for both from left to right and from right to left directions of the username. For the left side, the initial value of s is 0, while e is equal to 3 and it is increased by 1 at each iteration of the loop that breaks once the e gets equal to the length of the username. On the contrarily, for the right side, the initial value of

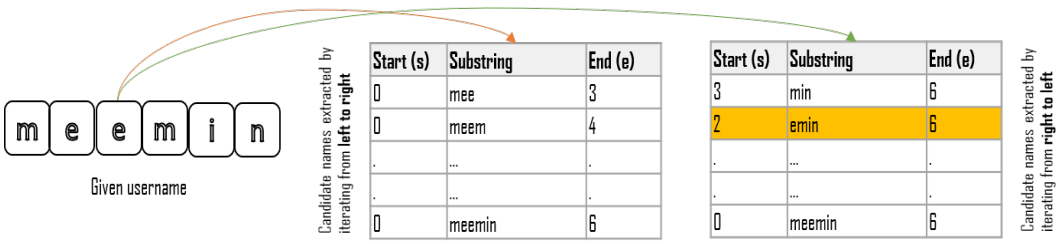


Fig. 2: Extraction of a set of candidate names from a given username by iterating on its both sides.

the e is equal to the length of username, while the s is equal to $e-3$ and it is decreased by 1 at each iteration of the loop that breaks once the s gets equal to 0. Using this method, possible candidates which have at least a length of 3 are extracted from a given username. Notice that candidates are restricted to have at least a length of 3 due to the fact that Turkish person names often have a greater length than 3. A simple visualization showing how this method creates substrings is depicted in Figure 2.

Note that this method can not catch a name (e.g., baturalp) that is located in the middle of the given string (e.g., fbbaturalpfb) and it needs to be improved for such an additional task. However, in this study, it is employed just as depicted in Figure 2, since users often locate their names at the head and tail sides of their usernames. Another reason behind this way of employing is that searching for other possible candidates will increase the computation cost of the method.

4.2. Longest common substring

Differently from the previously introduced substring extraction method, this algorithm is devoted to finding the longest common substring (LCS) between two given strings. For instance, The longest common substring of the strings "XYXYZ", "YXYZX", and "XYZYX" is string "XYZ" of length 3. Two main approaches used to implement the LCS algorithm are generalized suffix tree and dynamic proramming. In this study, the dynamic programming approach is used to implement LCS which requires $\Theta(N * K)$ time, where K and S represent the number of strings and total lengths of those strings respectively. The reader is advised to [45] for more details about the LCS algorithm.

In this study, LCS is used to detect whether substrings of two given usernames are concatenated in reverse order (e.g., aliosman $\rightarrow$ osmanali). To achieve this task, firstly LCS between two unequal (i.e., different) usernames is obtained, and then it is removed from both of the usernames. If the retaining substrings are equal it is considered that they actually contain the same substrings in reverse order.

4.3. Inference mechanisms

This study employs simple inference mechanisms to investigate how accurate an adversary can infer the username and gender of a user once he/she has access to even a very small part of the OSN data like in the leaked one [20], [42].

4.3..1 Gender inference mechanism

In this study, two different but quite similar approaches are employed to infer the gender of users. These approaches completely rely on

TABLE 1: Incomplete lists of Turkish person names contained within the offline and popularity-based lexicons used in this study

Dictionary/lexicon of popularity			Offline lexicon	
Name	Male	Female	Name	Target
Hasan	181,880	2,519	Kezban	Female
Yusuf	126,701	2,608	Ali	Male
Ali	297,092	4,786	Kibar	Unisex
Emine	1,879	80,395	Aycan	Unisex
Buket	205	5,879	Murat	Male
Pakize	71	2,660	Naciye	Female

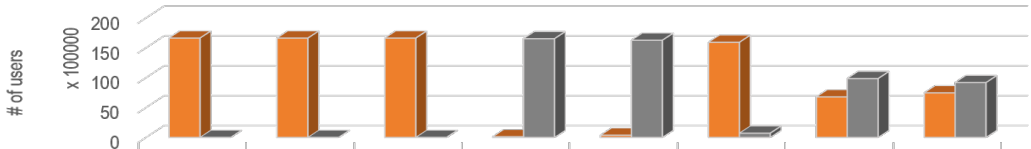
the dictionary/lexicon of person names in which each record is stored with the target gender or frequency of the name among both male and female users. The first approach in this study is referred to as the popularity-based approach [13], [16] which predicts the gender of a user based on a given name's popularity (i.e., frequency) among male and female users in the network. In this approach, the dictionary is created from the OSN data that the user at hand member of. On the other hand, the second approach referred to as offline lexicon-based approach [13], [16], [46] uses a lexicon of Turkish person names (see Section 3.2) and tries to predict a given user's gender by taking the matched record's target gender as its prediction.

Table 1 presents unordered and incomplete lists of dictionaries/lexicons used in both approaches. As seen in Table 1, popularity lexicon stores names along with their total observed frequencies within the first names of both male and female users. For instance, the first name "Hasan" is observed in the first names of 181,880 male and 2,519 female users respectively. As such, any user's gender with the first name "Hasan" will be predicted (i.e., inferred) to be a male since the number of male users using the same first name is higher than the number of female users. On the other hand, offline lexicon stores each name along with its target gender information. As such, this approach basically relies just on finding a match of queried first names with the lexicon. For instance, any user's gender with the first name "Ali" will be predicted to be male.

Note that the popularity-based lexicon is created by using the 16,624,540 first names (see Figure 3) of users within the leaked data, while offline lexicon (see Section 3.2) is taken from another project from the Internet. One should keep in mind that in the offline lexicon approach, if the queried first name does not have a match with the lexicon or its target gender is unisex even though it has a match, it is marked as unpredicted. On the contrarily, in the popularity-based approach, a user's gender is marked as unpredicted once his/her first name does not have a match with the lexicon. Besides, it is also marked as unpredicted again even though it has a match with the lexicon. Let the x and y be the number of male and female users within the lexicon such that they have the same first name as the queried name. The case comes into view under the following circumstances:

- if $x = y$,
- if $x = 1$ and $y = 0$,
- if $y = 1$ and $x = 0$,

This is because the popularity lexicon/dictionary is created by using the OSN data at hand and



	Phone Number	First Name	Surname	Email	Birth Date	Gender	Hometown	Place Liven-In
Disclosing	16624551	16624540	16624522	103867	365381	15915712	6780212	7454515
Undisclosing	0	11	29	16520684	16259170	708839	9844339	9170036

Fig. 3: The numbers of users w.r.t. their disclosing status of different attributes.

the last two conditions given above mean that the inference mechanism will make its prediction using the queried user's gender information. The unclassified (i.e., unpredicted) names in both approaches are not included in the confusion matrices, but they are taken into account when computing the overall accuracies of the approaches. Note that the overall accuracy is obtained dividing the number of users whose gender attribute is correctly predicted by the total number of users at hand. Further details are given in Section 5.2.1

4.3..2 Username inference mechanism

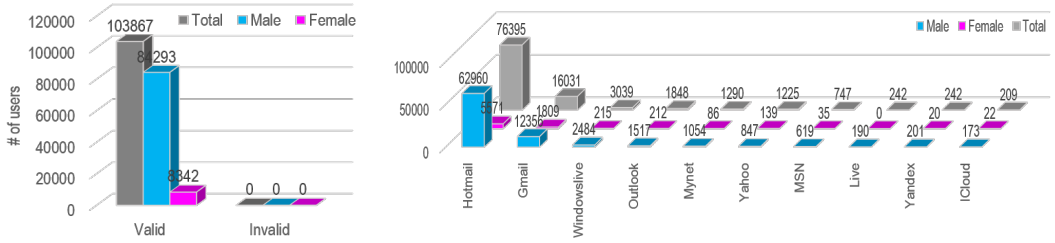
This is a very simple mechanism that investigates changes between two given usernames and mainly depends on simple string processing methods like LCS, split, replace, and so on. Further details on the employment of this mechanism are given in Section 5.2.2.

5. Results

This section presents the results of the statistical analysis of Facebook data along with the results of inference attacks devoted to detecting gender and email username attributes of users. The data is stored in a relational database with the help of the MySQL Workbench Database Management System (DBMS). Methods described in Section 4 are used along with the DBMS's aggregation functions to perform the analysis. Note that not all of the charts are given to save space, but they are also available (https://github.com/ocbn/fbleak) online to enable the readers to view them in high resolution as well.

5.1. Results of Statistical Analysis

As a first step of the experiments, a statistical analysis is performed to inspect sharing behavior of Turkish Facebook users. For this purpose, a simple table-where Facebook ID is configured to be the primary key-is designed to store the disclosed data in the DBMS. Then Facebook data moved to the relational database and it is detected that there exist a total of 19,638,818 user records from which 422,283 are duplicates and 2,591,984 are of non-Turkish users. As such, eliminating these approximately 3,01M (422,283 + 2,591,984) records has resulted in a total of 16,624,551 records of users in the designed relational database. Next, several aggregation functions are used to investigate which attributes are disclosed more frequently by users, how users behave while they selecting a username, which cities host the majority of users, and so on.



(a) (b)

Fig. 4: Analysis of email adresses: distribution of (a) valid and invalid mails w.r.t. gender, (b) emails w.r.t. email service providers.

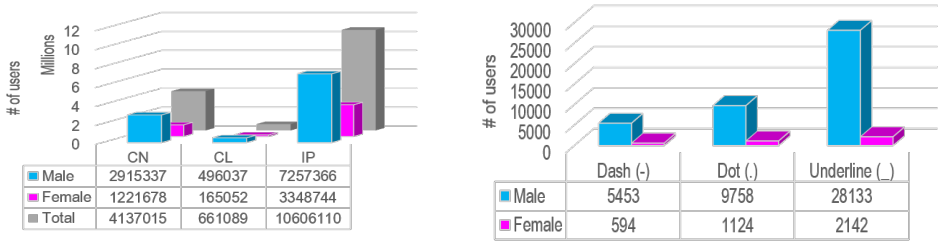
5.1..1 Publicity of attributes

As seen in Figure 3, email, and date of birth attributes are the least shared attributes by users, while the majority of users disclose their first name, surname, and gender attributes. Interestingly, all of the users shared their phone numbers within their accounts. To have a deeper look at attributes, each attribute is also investigated considering disclosing status, behavior, and so on. With this purpose, an analysis on the first name and gender attributes showed that among users disclosing their gender, 11,108,570 of those are male, while 4,807,142 of those are female. On the other hand, 640,409 male and 273,676 female users have a middle name. Besides, 66,587 male and 32,374 female users preferred to use a prefix with different formats such as “TC.”, “T.C”, “T.C.”, and “TC” in their display names. The most frequently used three first names of male users include Mehmet, Mustafa, and Ahmet, while those are Fatma, Ayse, and Emine for female users.

5.1..2 Email and date of birth

Next, another analysis is performed on email and date of birth attributes. This analysis showed that none of 365,381 (see Figure 3) users disclosed the exact date of birth which can be used to infer age attributes. Instead, users disclosed their date of the birth attribute in a way such that it only includes day and month information like May 06.

On the other hand, email addresses of 103,867 users are analyzed with regular expressions to investigate whether they are valid or invalid alongside which email service providers are mostly preferred by users. Results of these experiments are depicted in Figure 4 which shows that all of the disclosed email addresses are valid and the majority of users disclosing mail addresses are males, while the number of users disclosing a valid email address is 103,667, but not disclosing gender attribute is 11,232 (see Figure 4(a)). On the other hand, the distribution of email service providers w.r.t. male and female users is quite similar. Users take their email addresses mostly from Microsoft’s Hotmail, and Google’s Gmail services. Notice that service names are detected by splitting the email address by @ character and removing other strings - mostly concatenated by a dot - from the retaining string.



(a) (b)

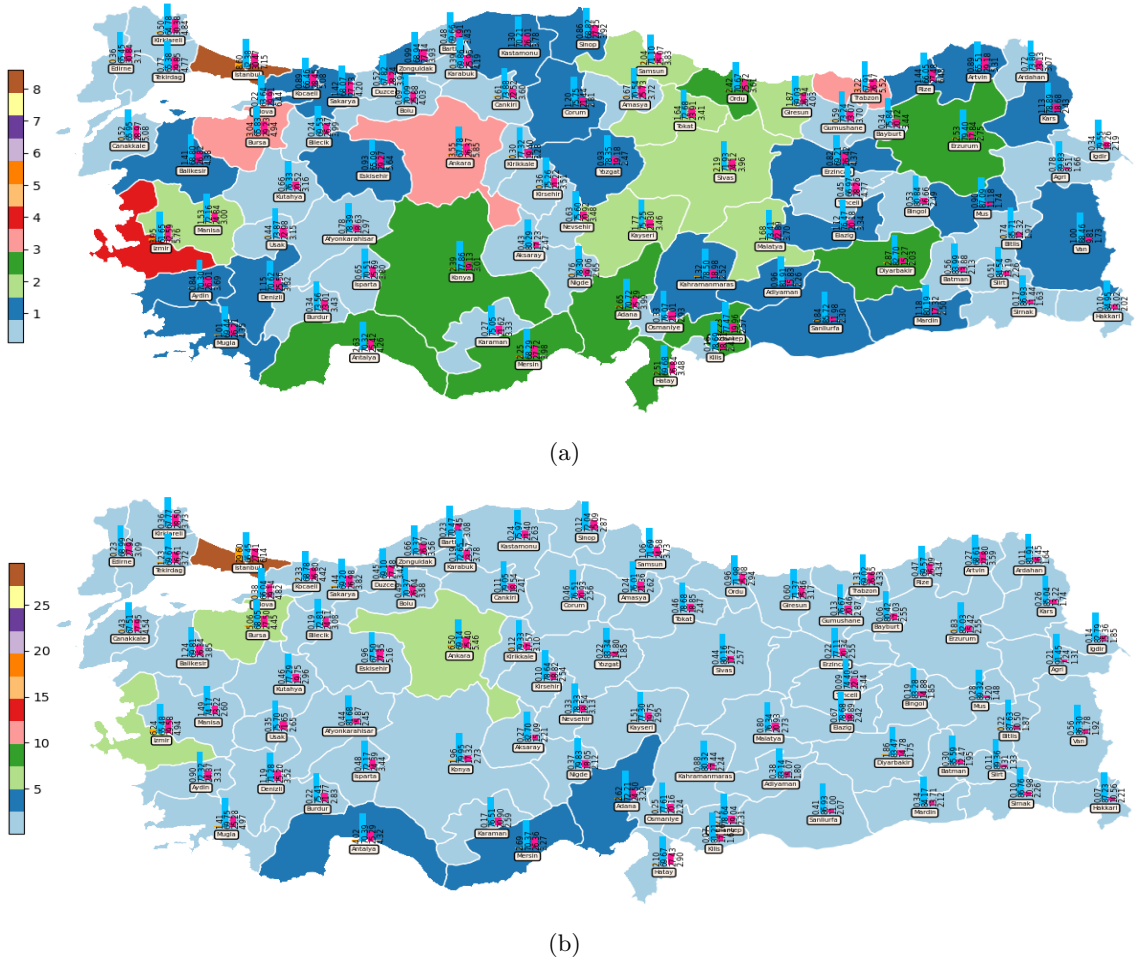
Fig. 5: Distribution of usernames w.r.t. gender on: (a) Facebook accounts, (b) email accounts.

5.1..3 Usernames

Afterward, Facebook and email usernames of users are analyzed to investigate the behavior of users while selecting/creating a username for their new accounts on Facebook and email service providers. The results of this analysis are depicted in Figure 5. In Figure 5(a), Facebook usernames of all users disclosing their gender (i.e., 15,915,712) investigated and found that 25.99 percent (i.e., 4,137,015) of users selected completely numeric usernames (i.e., CN) which are default ones suggested by Facebook. On the other hand, 4.15 percent (i.e., 661,089) of users selected usernames completely comprised of letters (i.e., CL), while 66.63 percent (i.e., 10,606,110) of users selected usernames including at least one punctuation mark (i.e., IP). Notice that all Facebook usernames including punctuation are created just by using the dot mark. Similar analysis on email usernames shows that 50,241 of 103,867 users disclosing their mail addresses created their usernames by using at least one punctuation mark. In these usernames only dash, dot, and underline marks are preferred with a percent of 13.16 (i.e., 6,614), 24.07 (i.e., 12,095), and 66.47 (i.e., 33,397) respectively. The results of the additional analysis - performed just on 92,635 users disclosing both email and gender attributes - depicted in Figure 5(b) show that the most preferred punctuation mark in email usernames is underline. On the other hand, the number of email usernames containing at least one punctuation mark is 45,487. The number of completely numerical usernames is 13 for males and 4 for females while the number of names consisting entirely of letters is 22,279 for males and 2,691 for females.

5.1..4 Geographical distribution of users

In the next step, an additional analysis is performed so as to extract the number of male, female, and total users w.r.t. their hometowns and places liven-in. In this phase, Turkey's geospatial data (see Section 3) is converted into a map with the help of geopandas (https://geopandas.org/) Python package. Results of this analysis are depicted in Figure 6, where an additional bar chart is inserted on each city's geographical center to also show the percent of users in that city across the country's population (at the most left and colored in orange), percent of male users across the city's population (colored in blue), percent of female users across the city's population (colored in pink), and percent of users not disclosing gender attribute across the city's population (at the most right and colored in gray) respectively.



(a)

(b)

Fig. 6: Distribution of user places considering: (a) gender and hometown, (b) gender and place lived-in.

As seen from Figure 6(a), the majority of users disclosed that they are from Istanbul, Ankara, Izmir, Bursa, Trabzon, and so on. The number of male users is higher than the number of female users in all cities. There also exists a heterogeneous distribution w.r.t. the total number of users across cities. What is more, the number of female users is often lower especially in the southeastward part of the country compared to the other parts.

As seen in Figure 6(b), on the other hand, this heterogeneous distribution is turned into being more homogenous when considering users' places lived-in. Again the number of male users is higher than the number of female users in all cities. The most crowded places of the country are Istanbul, Ankara, Izmir, Bursa, Antalya, Mersin, Adana respectively. These results show that most of the people living in big cities came from small cities of the country for various reasons. For instance, percent of the population in Istanbul is 8.60 when considering users from that city, while it increases to 29.60 when considering users living there. Majority of those users are from Trabzon (3.22 → 1.31), Sivas (2.19 → 0.44), Erzurum (2.53 → 0.83), Diyarbakir (2.87 → 1.86), Konya (2.39

→ 1.96), Gaziantep (2.23 → 2.12), and so on respectively. Note that the values in parentheses show the percentage of users from that city and the percentage of users who live in that city, respectively.

5.1..5 Geographical distribution of privacy-unaware users

Similar to the previous step, another but the last analysis is performed to inspect where the majority of privacy-unaware users are from and live-in. The same steps of the previous analysis are performed in this analysis, but differently, users who disclose all attributes are included in the evaluation.

Taking such a subset of users showed that 24,178 users disclosed all of the attributes, but 78 of them disclosed their hometowns that are out of Turkey. The geographical distribution of users within this subset is depicted in Figure 7, where an additional bar chart is inserted on each city's geographical center to also show the percent of unaware users across the total population of that city (at the most left and colored in orange), percent of male uses among unaware users in that city (colored in blue), and percent of female users among unaware users in that city (at the most right and colored in pink), respectively.

As seen from Figure 7(a), considering the hometown of users shows that Rize, Cankiri, Bilecik, and Kirsehir are cities where the most privacy unaware users are from with a percent of 0.26, 0.25, 0.24, and 0.24 respectively. On the other hand, Mus and Hakkari are cities where the least privacy unaware users are from. Interestingly, almost all of the privacy unaware users who are from the cities located in the country's southeast region are male.

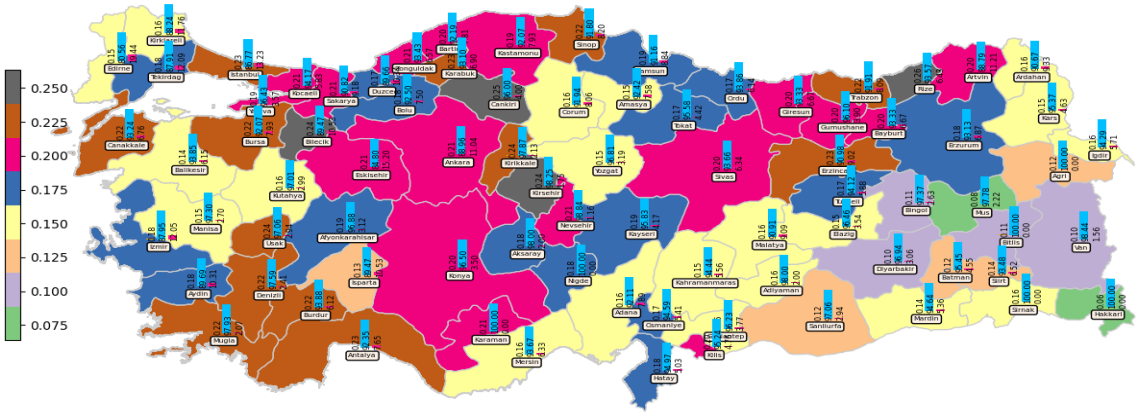
As seen in Figure 7(b), on the other hand, considering cities in which the most privacy unaware users live shows that cities with the highest number of unaware users compared to the population turned in to be Antalya, Usak, and Mugla respectively. On the contrarily, cities with the lowest unaware users are Bingol, Bitlis, Mus, Agri, and Igdir respectively. Besides, as observed in the previous step of the analysis, all unaware users are males who are living in the cities often located in the southeast region of the country. Sanliurfa, Mardin, Siirt, Hakkari, Van, Agri, Igdir, and Kars are to list a few examples. Besides, users who are at the secondary level with respect to the privacy risk live in Istanbul, Bursa, Denizli, Sinop, and Rize.

5.2. Results of Inference Attacks

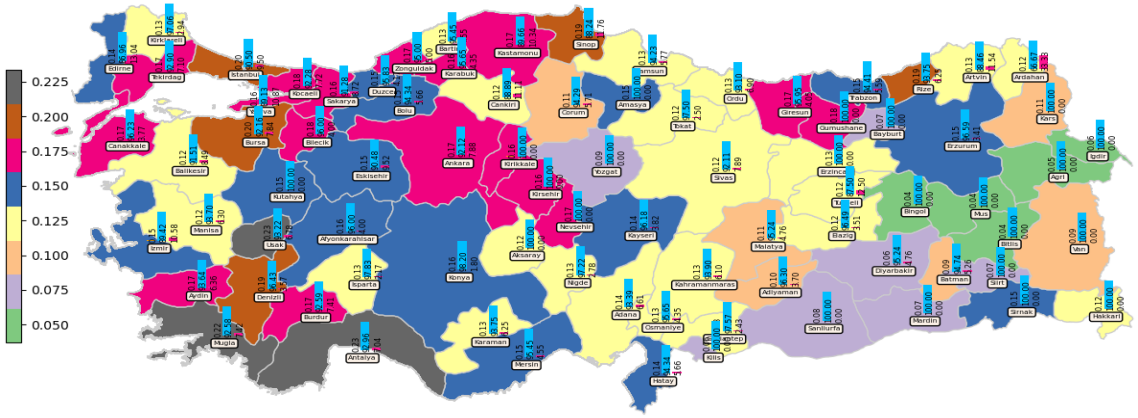
The richer the OSN data, the more inference attacks can be performed successfully. The leaked data at hand do not include connections, liked pages, wall activities, and other information, but just contains eight attributes (see Figure 3) of users. As the leaked data is incomplete (i.e., just includes a few profile attributes), it is not possible to perform a lot of different attacks, but there is still a possibility for gender and usernames. Therefore, in this study, two simple attacks are performed to infer gender attributes and usernames of users. This section presents the results of these inference tasks in the following sub-headings respectively.

5.2..1 Inferring gender

To infer the gender of users, lexicon-based and popularity-based approaches (see Section 4) are employed on both display names and usernames of users. If inference relies on display names, "TC"



(a)

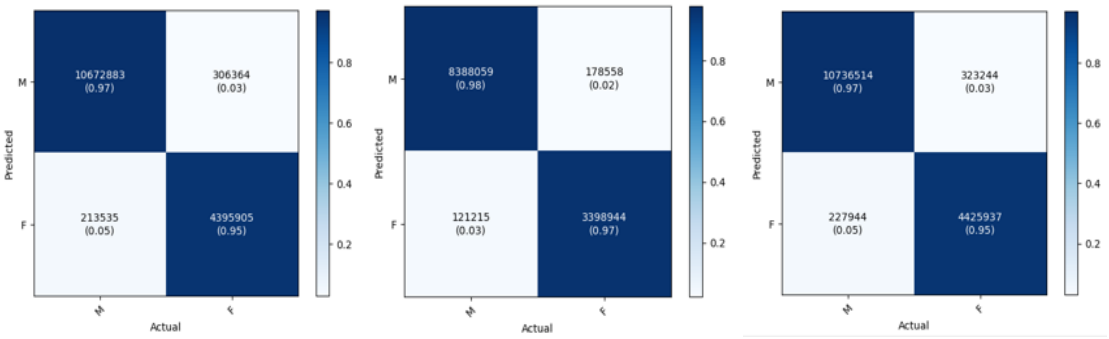


(b)

Fig. 7: Distribution of privacy-unaware users considering: (a) gender and hometown, (b) gender and place lived-in.

prefix is removed from the display name of a user, if exist. Then the display name is splitted with the help of the white spaces and the first token is accepted as the first name of user at hand. On the other hand, if inference relies on usernames, a set of candidate names are extracted using the method introduced in Section 4.1.

In the first step of display name based inference, the popularity-based approach (see Section 4) is employed which depends on a basic idea that the gender attribute of a user is predicted by considering the frequency of his/her first name among other users in the OSN. In this phase, it is detected that the value of gender attribute is empty for 28,981 of 15,915,712 users who disclose gender attributes. Excluding these users resulted in a user set including 15,886,731 users whose gender is known. Using this user set, firstly, gender inference is performed using the popularity approach on display names. As seen from Figure 8, 15,068,788 users's gender is correctly predicted, while 519,899 users' gender is incorrectly predicted. On the other hand, this simple inference mechanism is unable to predict the gender of 298,044 users. Notice that these unclassified users are not included in the confusion



(a) Overall acc: 0.948 (b) Overall acc: 0.741 (c) Overall acc: 0.954

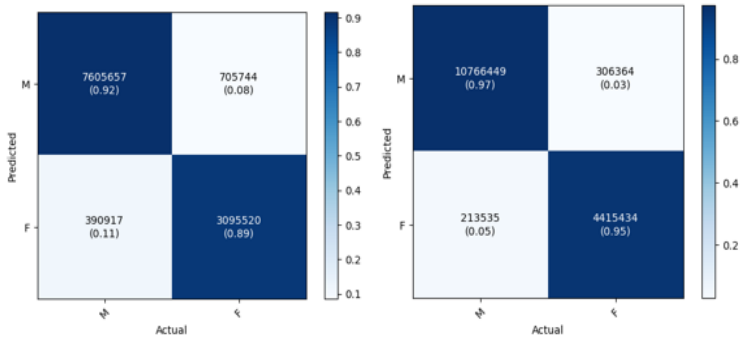
Fig. 8: Confusion matrices for the gender inference task relying on users' first names: (a) using popularity-based approach, (b) using offline lexicon approach, and (c) hybrid of the popularity- and offline lexicon-based approaches. Notice that unclassified users are not included in the confusion matrices.

matrix depicted in Figure 8(a), but computing the prediction accuracy considering both incorrectly classified and unclassified users shows that the overall accuracy of this first step is 0.948.

Secondly, the inference task is performed by using an offline lexicon described in Section 3.2. In this phase, the simple lexicon-based approach (see Section 4) is employed again the first names of users. The same preprocessing steps (removing prefix etc.,) are again applied on display names and a user's gender is considered to be correctly inferred only once his/her first name has a match in the lexicon and his/her gender is the same as the matched name's target gender. As seen from Figure 8(b), this way of inference has resulted in inferring 11,787,003 and 299,733 users' gender correctly and incorrectly respectively. On the other hand, 2,029,712 users' gender attribute is not classified due to their names matched a unisex name within the lexicon. Again considering the incorrectly and unclassified users show that this mechanism achieves an overall accuracy of 0.741.

In the third step, the gender inference is performed by using a hybrid of the previous two approaches. In this phase, it is tried to infer a user's gender by using his/her first name's popularity and if unable to classify it, it is searched within the lexicon of first names. As seen from Figure 8(c), using this combined way of inference mechanism classified 15,162,451 users' gender correctly, while it classified 551,188 users' gender incorrectly. On the other hand, the number of users whose gender attribute is not inferred is 173,092 and the overall accuracy is 0.954 which is higher compared to the accuracies of single uses of the combined approaches.

In addition to the inference experiments relying on the first names of users, an experiment is also performed by considering users' Facebook usernames to investigate how an inference mechanism based on usernames is accurate when it is not possible to use a user's first name. In this phase, the Facebook username is splitted by a punctuation mark, if it contains punctuation. Then each token of username is searched within the dictionary for the popularity of names. If any part has a match, the user's gender is inferred based on the frequency of the matched name's



(a) Overall acc: 0.908 (b) Overall acc: 0.955

Fig. 9: Confusion matrices for the gender inference task relying on the popularity of: (a) username, (b) first name and username. Notice that unclassified users are not included in confusion matrices.

popularity between male and female users. On the other hand, if the username does not contain any punctuation mark, candidate sub-strings (see Section 4.1) are used to infer the user's gender. If any of the candidate substrings has a match within the dictionary of the popularity of names, the user's gender is again inferred using the same way of the previous case. As seen from Figure 9(a), using this way of inference mechanism, 10,701,177 out of 11,797,838 users' gender is correctly inferred and overall accuracy of 0.908 is obtained. Notice that users who select completely numeric usernames are left out in this experiment.

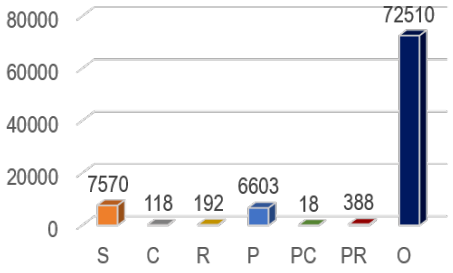


Fig. 10: Distribution of cases and changes made by users on their Facebook usernames when selecting/creating a new username for their email addresses.

Finally, gender inference is performed by using both the first name and Facebook username of user at hand. In this mechanism, it is tried to infer the gender of a user based on the popularity of his/her first name and if not classified, his/her username additionally is used to predict his/her gender. Notice that in this mechanism the additional inference step based on username is only applicable when the username is not comprised of completely numeric values. As seen from Figure 9(b), this two-step inference mechanism achieves the best accuracy among all of the mechanisms employed so far. It obtains an accuracy of 0.955 and correctly predicts the gender attribute of 15,181,883 users.

All of these results show that using an offline lexicon for gender inference is not a good choice while using the popularity of users' first names within the network helps a lot more to make highly accurate gender inference. Even though the best accuracy (i.e., 0.955) is obtained by using both the first names and usernames, the hybrid approach employed on the first names achieves a slightly different result which is an accuracy of 0.944.

5.2..2 Inferring username

As an additional attack scenario, it is also investigated whether it is possible to infer the Facebook username of a user by using his/her email address, and vice versa. For this purpose, 87,399 users who select Facebook usernames not comprised of completely numerical values and disclose their email addresses are inspected. The aim is that what are the changes that users prefer while selecting their usernames on Facebook and any of the email service providers.

The considered cases and changes are without any change (S), capitalization (C), replacing words (R), using at least one punctuation (P), P along with R (PR), P along with C (PC), and other (O) possible ones like adding suffix, prefix, etc. As seen from Figure 10, 7,570 users use completely the same (i.e., without any change) username for both their Facebook and email accounts. On the other hand, 118 users just capitalize the first character (e.g., "m.emin" → "M.emin") and 192 users replace words in reverse order (i.e., "ali.kemal" → "kemal.ali"), while 6,603 users just add at least one punctuation (e.g., "memin" → "m.emin_") to their newly created usernames. Notice that to detect username in which words are replaced in reverse order, the longest common substring algorithm (see Section 4.2) is used.

The results given in Figure 10 show that it is possible to infer the Facebook username of a user with a percent of 8.66 when his/her email username is given. This is also true for the case of inferring the email username of a user when his/her Facebook username is given. This percent can be increased up to 8.79 just by capitalizing the first character of the given username. To take one step further, inference still can be performed with a percent of up to 16.35 if the given username just contains punctuation changes.

6. Discussion

In this study, an analysis of the case of Turkish users on recently leaked Facebook data is performed along with two possible inference attacks. Analyzing sharing behavior on this subset of Turkish users' data shows that users often hesitate to disclose their email address and exact date of birth attributes. On the contrary, all of the users in this subset interestingly disclosed their phone numbers. Keeping email and exact date of birth attributes private can be interpreted as Turkish Facebook users are often aware of privacy implications of sharing their personal data on OSNs. However, they are still at great risk of privacy since they shared their phone numbers that can be used for several privacy and security implications. For instance, a user disclosing his/her phone number might get a scam from an adversary claiming to be someone the user knows. And in this way, the adversary may want the user to buy gift cards or wire money.

Another result of the analysis is that the number of male users is higher than the number of female users which means that males use Facebook more than females. Interpreting the geographical distribution of users based on their hometown and lived-in places shows that users mostly live in big cities like Istanbul, Ankara, and Izmir. On the other hand, the most privacy unaware users often live in big (e.g., Istanbul, Bursa) and coastal cities (e.g., Antalya, Mugla) of the country. The privacy unaware users disclose their all attributes highly because they are not aware of possible risks or share their information consciously. On the other hand, the region with the lowest Facebook users in the country is the

southeast region. Additionally, the number of female users living in this region is very low compare to the number of male users. Interestingly, almost all of the privacy unaware users in the region are male highly because female users may not actively use Facebook, not have a consistent internet connection, aware of the possible risk of privacy, or unaware of risks but just have an account without sharing anything among many other reasons.

Even though the leaked data has a very low number of attributes there still exists a possibility of performing inference attacks. As such, two simple inference attacks are performed to indicate this possibility. In the gender inference task, using an offline lexicon provides the worst results since the lexicon may be outdated and include unisex names which make predicting gender more challenging. On the other hand, using the popularity of any first name within the network help a lot more to achieve such a task. If the first name of a user is not available his/her username can also be used to infer his/her gender.

This is because OSN data would be more up-to-date compared to an offline lexicon. Most interestingly it is possible to infer the gender of a user just by using the first name of other users within the network. In other words, any OSN user can help an adversary to infer another user's gender just by disclosing his/her gender and real first name. Besides, users are akin to use similar usernames for their accounts by making several changes such as adding punctuation, making capitalization, and so on. A detailed investigation on these changes shows that an adversary has a chance with a percent of 8.66 to correctly infer an email username once a Facebook username is given, and vice versa. These results of inference attacks show that selected/created usernames for both Facebook and email accounts should not include information redundancies about the users or must not give clues about users' personalities. For instance, selecting a completely numeric username prevent possible inference attacks aimed to learn gender attribute.

Consider that these analysis and attack scenarios performed on a very small part of Facebook data which does not include users' social connection, liked pages, wall activities, and so on. This means that an adversary is capable of violating an OSN user's privacy even though he/she has access to a very small part of the OSN data. Therefore, the risk for OSN users will be much more especially when the adversary has access to a larger portion of OSN.

7. Conclusion

In this study, an exploratory analysis of recently leaked data of Turkish users is performed to inspect sharing behavior of users and perform possible inference attacks. Based on the results, it is concluded that male users are more active than female users on Facebook. Disclosing phone numbers can still be put users at risk even they do not disclose other personal attributes like email and exact date of birth.

In addition, even users keep their attributes private their information can still be inferred using various techniques and the accuracy of an inference attack will be much higher depending on the completeness of OSN data. Even though using very simple techniques, the gender attribute of users inferred with an accuracy of up to 0.95 which proves this reality. As such, it is also concluded that users should not disclose their personal information in OSNs or at least should use privacy settings that restrict the access grants of other users to their information.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

References

[1] K. Berger, J. Klier, M. Klier, and F. Probst. "A review of information systems research on online social networks", Communications of the association for Information Systems, Vol.35, pp. 145-172, September 2014.

[2] Y. A. Modi and I. S. Gandhi. "Internet sociology: Impact of Facebook addiction on the lifestyle and other recreational activities of the Indian youth", Proceedings of the The International Conferences on Socio-Cultural, Anthropology, Criminology and International Relations, Jakarta, Indonesia, pp. 1-4, 14-16 October 2013.

[3] Anoynmous, "The number of worldwide social network users", Statistica Research Department, [Online], Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, 2021.

[4] J. Zhang and S. Y. Philip. "Broad Learning Through Fusions", Springer International Publishing, Switzerland , 2019

[5] K. Alemerien. "Usable Security and Privacy on Online Social Networks: Tools, Approaches, Studies, and Future Trends", International Journal of Software Innovation (IJSI), Vol.9, No.2, pp. 35-68, 2021.

[6] Y. Li, Y. Peng, W. Ji, Z. Zhang, and Q. Xu. "User identification based on display names across online social networks", IEEE Access, Vol.5, pp. 17342-17353, 25 August 2017.

[7] V. Cosenza. "World map of social networks", Vincos Blog, [Online], Available: https://vincos.it/world-map-of-social-networks/, 2021

[8] D. Gayo Avello. "All liaisons are dangerous when all your friends are known to us", Proceedings of the ACM Conference on Hypertext and hypermedia, Eindhoven, Netherlands, pp. 171-180, 6-9 June 2011.

[9] M. Kiranmayi and N. Maheswari. "A review on privacy preservation of social networks using graphs", Journal of Applied Security Research, Vol.16, No.2, pp. 190-223, 23 April 2020.

[10] O. Coban, A. Inan, and S. A. Ozel. "Your Username Can Give You Away: Matching Turkish OSN Users with Usernames", International Journal of Information Security Science, Vol.10, pp. 1-15, March 2021.

[11] O. Coban, A. Inan, and S. A. Ozel. "Privacy Risk Analysis for Facebook Users", Proceedings of the IEEE Signal Processing and Communications Applications Conference, Gaziantep, Turkey, pp. 1-4, 5-7 October 2020.

[12] D. Choi, Y. Lee, S. Kim, and P. Kang. "Private attribute inference from Facebook's public text metadata: a case study of Korean users", Industrial Management & Data Systems, Vol.117, pp. 1687-1706, September 2017.

[13] O. Coban, A. Inan, and S. A. Ozel. "Facebook Tells Me Your Gender: An Exploratory Study of Gender Prediction for Turkish Facebook Users", Transactions on Asian and Low-Resource Language Information Processing, Vol.20, No.4, pp. 1-38, May 2021.

[14] Y. Kilic and A. Inan. "Implementing A Web Crawler With An Attacker Perspective On A Professional Purpose Online Social Network", Proceedings of the International Conference on All Aspects of Cyber Security, Adana, Turkey, pp. 27-32, 25 October 2019.

[15] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. "Inferring private information using social network data", Proceedings of the 18th international Conference on World wide web, Madrid, Spain, pp. 1145-1146, 20-24 April 2009.

[16] C. Tang, K. Ross, N. Saxena, and R. Chen. "What's in a name: A study of names, gender inference, and gender behavior in facebook", Proceedings of the International Conference on Database Systems for Advanced Applications, Hong Kong, pp. 344-356, 22-25 April 2011.

[17] C. Cadwalladr and E. Graham-Harrison. "Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach", The Guardian, [Online]. Available: https://www.theguardian.com/technology/2018/apr/08/facebook-to-contact-the-87-million-users- affected-by-data-breach, 2018.

[18] S. Nick. "Maker of popular quiz apps on Facebook exposed data of 120 million users", [Online]. Available: https://www.theverge.com/2018/6/28/17514822/facebook-data-leak-quiz-app-nametests-social-sweetheartexposed-user-info, 2018.

[19] J. Garside. "Twitter puts trillions of tweets up for sale to data miners", The Guardian. [Online]. Available: https://www.theguardian.com/technology/2015/mar/18/twitter-puts-trillions-tweets-for-sale-data-miners, 2015.

[20] D. Uberti. "Facebook Says Leak of 533 Million Users' Data Wasn't a Hack. Does it Matter?", The Wall Setreet Journal, [Online]. Available: https://www.wsj.com/articles/facebook-says-leak-of-533-million-users-data-wasnt-a-hack-does-it-matter-11617910106, 2021.

[21] O. Coban, A. Inan, and S. A. Ozel. "Towards the design and implementation of an OSN crawler: A case of Turkish Facebook users", International Journal of Information Security Science, Vol.9, pp. 76-93, June 2020.

[22] O. Coban, A. Inan, and S. A. Ozel. "Inverse document frequency-based sensitivity scoring for privacy analysis", Signal, Image and Video Processing, pp. 1-9, August 2021.

[23] O. Coban, A. Inan, and S. A. Ozel. "Fine-grained Kinship Detection for Facebook Users based on Wall Contents", Proceedings of the IEEE Innovations in Intelligent Systems and Applications Conference, Elazığ, Turkey, pp. 1-4, October 2021.

[24] O. Kulcu and T. Henkoglu. "Privacy in social networks: An analysis of Facebook", International Journal of Information Management, Vol.34, pp. 761-769, December 2014.

[25] E. Avllazagaj, E. Ayday, and A. E. Cicek. "Privacy-Related Consequences of Turkish Citizen Database Leak", Proccedings of the International Network for Economic Research Conference, Darmstadt, Germany, pp. 1-18, 8-10 June 2016.

[26] E. Kahya-Ozyirmidokuz. "Analyzing unstructured Facebook social network data through web text mining: A study of online shopping firms in Turkey", Information Development, Vol.32, pp. 70-80, January 2016.

[27] O. Coban, S. A. Ozel, and A. Inan. "Deep Learning-based Sentiment Analysis of Facebook Data: The Case of Turkish Users", The Computer Journal, Vol.64, pp. 473-499, January 2021.

[28] O. Coban, B. Ozyer, and G. T. Ozyer. "Sentiment analysis for Turkish Twitter feeds", Proceedings of the IEEE Signal Processing and Communications Applications Conference, Malatya, Turkey, pp. 2388–2391, 16-19 May 2015.

[29] H.A. Shehu, M. H. Sharif, M. H. U. Sharif, R. Datta, S. Tokat, S. Uyaver, and R. A. Ramadan. "Deep Sentiment Analysis: A Case Study on Stemmed Turkish Twitter Data", IEEE Access, Vol.9, pp. 56836-56854, April 2021.

[30] H. Karayigit, C. I. Aci, and A. Akdagli. "Detecting abusive Instagram comments in Turkish using convolutional Neural network and machine learning methods", Expert Systems with Applications, Vol.174, pp. 1-15, July 2021.

[31] C. Coltekin. "A corpus of Turkish offensive language on social media", Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, pp. 6174-6184, 11-16 May 2020.

[32] S. A. Ozel, E. Sarac, S. Akdemir, and H. Aksu. "Detection of cyberbullying on social media messages in Turkish", Proceedings of the IEEE International Conference on Computer Science and Engineering, Antalya, Turkey, pp. 366-370, 5-8 October 2017.

[33] A. Bozyigit, S. Utku, and E. Nasibov. "Cyberbullying detection: Utilizing social media features", Expert Systems with Applications, Vol.179, pp. 1-12, October 2021.

[34] O. Ozdikis, P. Senkul, and H. Oguztuzun. "Semantic expansion of tweet contents for enhanced event detection in twitter", Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Turkey, pp. 20–24, 26-29 August 2012.

[35] D. Kucuk. "Sentiment, Stance, and Intent Detection in Turkish Tweets", In New Opportunities for Sentiment Analysis and Information Processing, IGI Global Inc., USA, 2021.

[36] M. Kaya, G. Fidan, and I. H. Toroslu. "Sentiment analysis of Turkish political news", Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Macau, China, pp. 174–180, 4-7 December 2012.

[37] M. Ciot, M. Sonderegger, and D. Ruths, D. "Gender inference of Twitter users in non-English contexts", Proceedings of the Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, pp. 136-1145, 18-21 October 2013.

[38] E. Sezerer, O. Polatbilek, and S. Tekir. "Gender prediction from Turkish tweets with neural networks", Proceedings of the IEEE Signal Processing and Communications Applications Conference, Sivas, Turkey, pp. 1-4, 24-26 April 2019.

[39] E. Sezerer, O. Polatbilek, and S. Tekir. "A Turkish Dataset for Gender Identification of Twitter Users", Proceedings of the Linguistic Annotation Workshop@ ACL, Florence, Italy, pp. 203–207, 1-2 August 2019.

[40] M. Talebi and C. Kose. "Identifying gender, age and education level by analyzing comments on Facebook", Proceedings of the IEEE Signal Processing and Communications Applications Conference, Haspolat, Turkey, pp. 1–4, 24-26 April 2013.

[41] O. Celik and A.F. Aslan. "Gender prediction from social media comments with artificial intelligence", Sakarya Universitesi Fen Bilimleri Enstitusu Dergisi, Vol.23, pp. 1256–1264, December 2019.

[42] J. Peters. "Personal data of 533 million Facebook users leaks online", The Verge, [Online]. Available: https://www.theverge.com/2021/4/4/22366822/facebook-personal-data-533-million-leaks-online-email-phone-numbers, 2021.

[43] I. Baskin. "A database of Turkish person names", Github, [Online]. Available: https://gist.github.com/ismailbaskin/1325813/9157dd8ced294a11218449d43bf9f772780f5d85

[44] Anonymous, "A database of cities and district names of Turkey", Github, [Online]. Available: https://gist.github.com/rainb3rry/6bbf945118362b1509adb46d95bca30c

[45] A. Amir, P. Charalampopoulos, S. P. Pissis, and J. Radoszewski. "Dynamic and internal longest common substring", Algorithmica, Vol.82, pp. 3707–3743, July 2020.

[46] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, and L. H. Ungar. "Personality, gender, and age in the language of social media: The open-vocabulary approach", PloS one, Vol.8, pp. 1-16, September 2013.