IJE TRANSACTIONS B: Applications Vol. 36, No. 02, (February 2023) 335-347

International Journal of Engineering

Journal Homepage: www.ije.ir

Gender Identification of Mobile Phone Users based on Internet Usage Pattern

F. Negaresh, M. Kaedi*, Z. Zojaji

Faculty of Computer Engineering, University of Isfahan, Azadi Sq., Hezarjarib St., Isfahan, Iran

PAPER INFO

Paper history:
Received 30 September 2022
Received in revised form 29 November 2022
Accepted 02 December 2022

Keywords:
Gender Identification
Internet Usage Pattern
User Modeling
Machine Learning
Feature Selection

ABSTRACT

Gender is an important aspect of a person's identity. In many applications, gender identification is useful for personalizing services and recommendations. On the other hand, many people today spend a lot of time on their mobile phones. Studies have shown that the way users interact with mobile phones is influenced by their gender. But the existing methods for identify the gender of mobile phone users are either not accurate enough or require sensors and specific user activities. In this paper, for the first time, the internet usage patterns are used to identify the gender of mobile phone users. To this end, the interaction data, and specially the internet usage patterns of a random sample of people are automatically recorded by an application installed on their mobile phones. Then, the gender identification is modeled using different machine learning classification methods. The evaluations showed that the internet features play an important role in recognizing the users gender. The linear support vector machine was the superior classifier with the accuracy of 85% and F-measure of 85%.

doi: *10.5829/ije.2023.36.02b.13*

NOMENCLATURE

FP	False Positive	TP	True Positive
FN	False Negative	TN	True Negative

1. INTRODUCTION

Nowadays, the benefits of having a smartphone are undeniable, and almost everyone uses it almost constantly. For this reason, these devices contain rich sources of information about users and are powerful tools for better understanding of the user [1]. But different people use mobile phones differently. Various factors such as age, level of education, job, personality characteristics, and gender affect the people's mobile phones usage and internet usage patterns. Actually, studies have shown that the way users interact with mobile phones is influenced by their gender [2-4]. In particular, the internet usage patterns in male and female users are not the same.

On the other hand, user gender identification can play an important role in personalizing e-commerce services. For example, from the marketing perspective, the analysis of preferences and target items of each gender provides effective marketing strategies and profitable offers for companies [5, 6]. If a mobile application can identify the users gender and then provides personalized services according to the users gender, the experience of people interacting with smartphones will be more enjoyable, which increases customer satisfaction and loyalty and, consequently, the profitability of businesses.

Due to the importance of automatic gender identification of mobile phone users, this field has been considered in several studies. However, some of the methods suffer from low accuracy [3], some require the use of special sensors [4] (e.g., accelerometers and gyroscopes), and some others necessitate specific user activities such as high-speed walking that are not applicable to users who have been sedentary.

The objective of this study is to use the internet usage pattern of mobile phone users to identify their gender. The hypothesis that has been examined in this study is that the use of users internet usage patterns, alone or along with other features, can lead to the accurate users gender identification. In this paper, we intend to propose

*Corresponding Author Institutional Email: kaedi@eng.ui.ac.ir (M. Kaedi) a method that predicts the users gender by investigating their Internet usage pattern, while it does not require special sensors on the mobile phone and does not depend on the specific user activities.

In the proposed method, a sample of 99 mobile phone users were randomly selected. Then, by installing an application on their mobile phone, their daily interactions with the mobile phone were automatically recorded. The obtained data were then analyzed and several classification models were built to predict the users gender. Finally, the results of gender classification for different scenarios were evaluated using standard evaluation criteria.

In the rest of the paper, in section 2, an overview of the related work is presented. Then, in section 3, the proposed method is discussed in details. Section 4 describes the results obtained, and section 5 presents the conclusions and future work.

2. LITERATURE REVIEW

In this section, the related work are reviewed in three categories.

The first category which is the most relevant to the current study includes studies that identifies the gender of mobile phone users. Jane and Kanhangad [4, 7] performed gender classification using user's gait information recorded by smartphone internal sensors. The authors recorded the user's gait data, including the data collected by accelerometer and gyroscope sensors, using an Android application. They then used gradients to extract several features from this data. Finally, by applying the decision tree learning algorithm to pre-processed data, they identified the user gender with 90% accuracy. In another study, Jane and Kanhangad [3] developed an approach to gender recognition in smartphones using user touch screen gestures. They used some classification algorithms to recognize the finger gestures pattern of - males and females. The K-nearest neighbor classification algorithm showed the highest accuracy in identifying user gender. In their study, Sarraute et al. [8] investigated whether the differences in smartphone usage between males and females are reflected in their call and SMS patterns. They were interested to find out the difference between the use of mobile phones in different age groups. Afterwards, they predicted the age and gender of users using several algorithms including Naïve Bayes, support vector machine (SVM), and logistic regression algorithms, which at the best resulted in 62.3% accuracy. Choi et al. [9] collected user data from messengers and social media platforms and investigated the wrodsets preferred by male and female users to predict their gender. They identified a set of representative wordsets for men and women and identified the gender of writers based on the presence of representative words in the texts. They used several measures, such as: term frequency–inverse document frequency (TF-IDF), mutual information and chi-square to calculate the similarity of the words in the text with the representative words. They examined Naive Bayes classifiers, logistic regression, and SVM with a linear kernel. They showed that SVM performed better than the others. Miguel-Hurtado et al. [10] presented an approach to identify the user gender based on touch gestures and keystroke dynamics. They collected touch data on both the horizontal and vertical axes, and the feature vectors were defined as the mean of the collected touch features. The dataset included 57 men and 59 women, and the evaluation was performed through 10-fold cross validation .

Despite the success of the studies of the first category, they suffer from significant limitations and weaknesses. Methods based on the user gait information pattern require specific user activity. They are also device-dependent due to their dependence on accelerometer and gyroscope sensors, and generalizing them to other devices with different versions of sensors is a challenging task. Also, the accuracy of these methods depends significantly on the walking speed. In addition, signals received from gyroscope and accelerometer sensors produce large volumes of data that cannot be easily be stored and processed by the mobile phone. Therefore, processing these signals and recognizing the gender accordingly may not be easily achievable in real-time. The large volume of data received for touch gestures-based methods is also an important challenge that makes it difficult to be used in a real-time gender recognition model. Also, accessing messages sent by mobile users can violate user privacy, and many users are reluctant to give permission to an application for accessing their messages. Unlike previous methods, using the Internet connection pattern presented in our study considers higher-level data, which, in addition to using a much lower volume of data than sensor signals, is not dependent on specific devices or specific sensors and preserves the user's privacy.

The second category includes studies on how females and males use the internet. Ramirez-Correa et al. [11] studied the pattern of internet usage between females and males and found that the pattern of internet usage is influenced by gender. In their study, the use of the internet by smartphone users was analyzed. The results show that males who have smartphones are more inclined to use mobile Internet. Su et al. [12] investigated the pattern of internet traffic volume and usage duration of men and women. By analyzing 204,352 mobile phone users from 34 countries, they found that men spend more time online than women and their use of the internet has a direct impact on their health, alcohol use, and smoking. In another study, Okazaki and Hirose [13] examined the impact of gender on the use of smartphones by Japanese men and women to search for travel information. They used an online questionnaire and collected 992 responses. They observed that women were more likely to use the mobile internet to search for travel information. In addition, the structural equation model showed that men are more likely to use mobile phone than traditional personal computers. Also, mobile internet usage is more popular among females than males. Even thought studies reviewed in this category investigate the difference in the volume and manner of internet usage between men and women, they have not provided a model for gender identification; they presented only the difference analysis.

In the third category, studies are introduced that identify the gender of users in social networks regardless of the device used. Vashisth and Meehan [14] categorized the gender of the Twitter users based on the text of their tweets. They first extracted features from tweets using natural language processing techniques. Then, common machine learning methods (e.g., logistic regression, SVM, and Naïve Bayes) were applied to the extracted features. The results showed that the traditional Bag of Words models could not produce accurate results in gender recognition; however, word embedding models can work better than several machine learning methods. Therefore, in their study, word embedding models have been introduced as the most efficient method in gender classification based on Twitter textual data. Abadin et al. [15] showed that the way that social media users use language can reflect their gender. They concluded that to identify the writer's gender, texts should be analyzed from both psycholinguistic and semantic aspects. After extracting various features, they used different classification methods such as Random Forest, AdaBoost, and LightGBM [16] to identify the gender. Finally, they reported the performance of each of the three classification methods using different feature categories, with and without feature selection. Results showed that using all types of features, regardless of the feature selection step, leads to the best results. Kowsari et al. [17] used ensemble deep learning to categorize the profiles of Twitter and Facebook users into male and female groups. They utilized Random Multimodel Deep Learning, which consists of several deep neural networks and a convolutional neural network in which the architecture and the number of nodes are generated randomly. Safara et al. [18] employed the Whale optimization algorithm as a metaheuristic method to find latent patterns of the text in combination with a two-hidden-layer neural network to identify the gender of the users. The Enron dataset, containing half a megabyte of e-mail texts from social network users, was adopted to evaluate their proposed method. studies reviewed in this category have identified gender on social media generally regardless of considering specific limitations of mobile phones such as limited processing power and privacy issues. These studies have mainly identified gender based on the content of posts shared on social networks and sometimes based on user profiles on social networks, which cannot be accessed on mobile phones without the user's privilege. Furthermore, the gender classification error is often relatively high in these methods, which threatens their usefulness in real-world applications .

3. PROPOSED METHOD

This study presents a gender identification system for mobile phone users based on user interactions, especially their internet usage pattern. The architecture of the proposed method shown in Figure 1 includes the following steps.

- Data acquisition from users interactions with mobile phones along with the users gender
- Data processing and feature extraction
- Feature selection to detect the features that are effective in identifying the gender of users
- Training the gender classification models by applying machine learning algorithms to the training data
- Applying the models to test data to evaluate their gender prediction accuracy

As it is shown in Figure 1, user-mobile interaction data is first collected by an Android application. After preprocessing, the features affected by the user gender are extracted from the raw data, and then, significant features are selected. After splitting the data into training and test subsets, modeling is performed on the training data. Finally, the test data is used to evaluate the performance of gender prediction models. The details of each step are described in the following sections.

3. 1. Data Acquisition In order to collect data, a random sample of 99 mobile phone users from different strata of the society was selected. The set included 47 men and 52 women. The age of these people was between 17 and 41 years, with the mean of 24.3, the median of 22, and the standard deviation of 5.8 years. These users were asked to install a specific Android application called Mobisense [19], designed to collect data on user-mobile phone interactions. Participants gradually joined in and installed the android application. The data acquisition process took eight months. After installation, the application does not require direct user intervention but implicitly records user behavioral data. This application records two types of data :

-User gender which is first asked directly from the user when installing the application. This data is used as the class labels in the model training process.

-Behavioral data such as internet usage, call, and SMS history, which is automatically collected by the application to be used as the inputs for classification models.

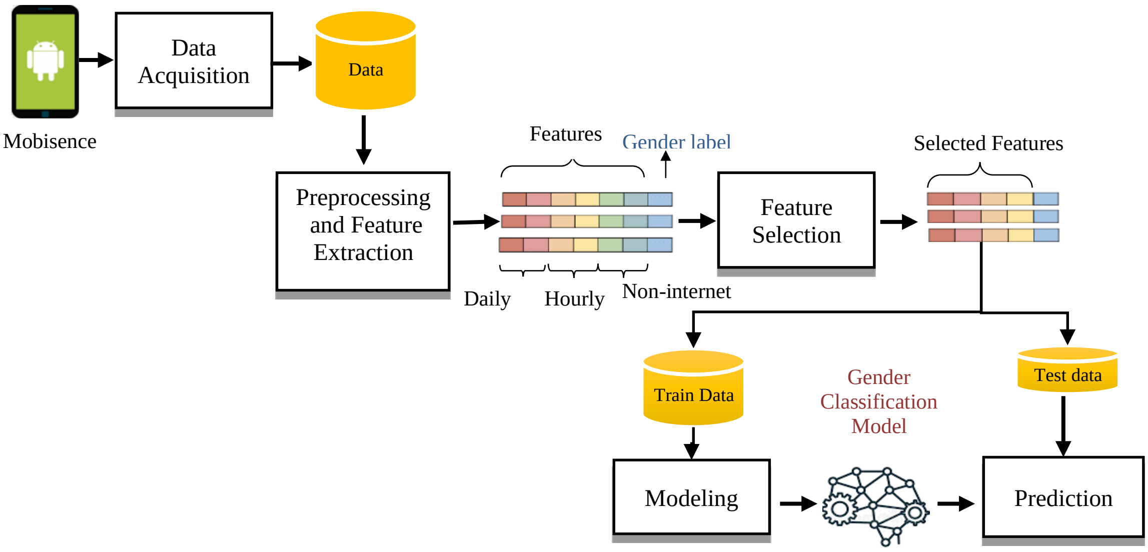

Figure 1. The architecture of the proposed method

In previous methods, the questionnaire has been used to collect data. However, data collected by questionnaire is influenced by the moods and situations of the participant at the time of completing the questionnaire. In addition, it is not possible to collect data continuously using a questionnaire. Unlike previous methods, the proposed method uses Android application to collect data implicitly and continuously, without being biased by the user's moods.

3. 2. Preprocessing

Some users were excluded from future analysis in the preprocessing step because insufficient data was collected from them. To be more precise, users whose data was collected for less than one day, did not use mobile internet, or used multi-user phones were excluded from the list. Finally, 60 users remained, including 28 men and 32 women. Therefore, the final data included the user gender and behavioral features of individuals whose mobile phone interactions were recorded within an acceptable time frame, who used the mobile internet during that period, and all interactions belonged to only one user of a specific gender.

3. 3. Feature Extraction

After data collection and data preprocessing, the data were stored as application logs in the SQL database to perform the feature extraction step. In this step, the significant features are extracted and calculated based on the raw data. Ninety-three features were extracted for each user, including cumulative and partial internet-based features and other features related to user interactions with the mobile phone (hereinafter referred to as "non-Internet-based" features).

3. 3. 1. Internet-based Features

As shown in Table 1, two categories of internet-based features were extracted in this step. The first category is cumulative Internet-based features that describe the overall Internet usage of user throughout the day. Cumulative features do not indicate the distribution of Internet usage at different hours of the day, while Internet usage at different hours can differentiate the gender of the user. Therefore, the detailed internet-based features is defined as the second category of features. This category includes 12 detailed features (reffered to as D1, …, D12) indicating the percentage of daily internet connection at two-hour intervals throughout the day.

TABLE 1. Internet-based features

Feature Category	Feature Name	Description
Cumulative	C1	The average number of times the mobile data network is turned on per day
	C2	The average number of times Wi-Fi is turned on per day
	C3	The average number of network connections per day
	C4	The average duration of network connection per day
Detailed	D1	What percentage of the user's daily connection was between 24 and 2 o'clock.
	D2	What percentage of the user's daily connection was between 2 and 4 o'clock.
	…	…
	D12	What percentage of the user's daily connection was between 22 and 24 o'clock.

The following equation shows how to calculate the D1:

$$D1=100\times(\text{Internet connection duration from 24 to 2 o'clock}) / (\text{Internet connection duration per day}) \quad (1)$$

Other features, including D2 to D12, are calculated in a similar way for their respective periods.

3. 3. 2. Non-internet-based Features

In addition to internet-based features, other behavioral features of user interactions with mobile phones have also been recorded and used. These features are called non-internet-based features in this study. They fall into 14 categories related to the following topics: call logs, hands-free usage logs, GPS usage logs, user profiles, applications installation, update, and uninstallation logs, airplane mode setting logs, Bluetooth usage logs, language change logs, battery level logs, power supply logs, sound and vibrate setting logs, touch screen turn on/off logs, SMS logs, and time zone change logs.

3. 4. Feature Selection

Some features may not be important in determining the gender of a mobile phone user; The presence of such features in the training phase can cause bias in the model or reduce the accuracy. Therefore, using a feature selection step before applying classification models can lead to the selection of useful information, thus increasing the accuracy of model. For feature selection, the random forest algorithm was employed. In this method, after creating a random forest, an importance factor is calculated for each feature according to its average ability to increase the pureness of the leaves over the trees of the forest. Then, the most important features are selected [20].

3. 5. Training

The set of features extracted for each user forms a feature vector that can be used to train a binary classifier. Classification is a supervised learning algorithm that, given a set of feature vectors and labels, models the input-output relationship and classifies new data based on the obtained model (male = 1and female =0). In this study, seven machine learning classification algorithms, including LSVM with Radial basis kernel, KNN, Naïve Bayes, CART [21] decision tree, random forest, and Adaboost [22, 23] have been used. The reason for choosing these algorithms was to use and compare the performance of different types of classifiers. Some of these algorithms are linear, and some others are nonlinear. Also, some are stochastic and some are deterministic. In addition, random forest and Adaboost are ensemble learning methods utilizing multiple learning algorithms to obtain better performance. In general, modeling the data by various algorithms gives a more comprehensive view of the problem at hand; therefore, a more accurate comparison and evaluation can be made. The set of candidate classifiers was selected the same as what was used by the related work so that the proposed method could be comparable to the literature. Hence, the effect of using internet-based features is determined purely.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed method, different classification algorithms were applied and evaluated through 10-fold cross-validation [24]. In this evaluation method, the set of all users was randomly divided into ten folds. In each iteration, a fold is considered as the test data, and the model was trained using the remaining folds. This process was repeated for each of the ten folds, and the final evaluation criteria were calculated by averaging the performance over all iterations. Moreover, different scenarios were designed to investigate the effect of different feature sets on the performance of the gender identification models. In the rest of this section, after introducing the evaluation criteria and hyperparameter settings, different feature selection scenarios are evaluated and compared.

4. 1. Evaluation Criteria

The evaluation criteria used in the current study include accuracy, precision, recall, and F-measure, which are expressed in Equations (2) to (5) based on the confusion matrix shown in Table 2. Male and female are considered as the positive and negative classes, respectively. Due to the trade-off between precision and recall, modifying the models to increase precision usually results in reducing the recall and vice versa. To balance these two criteria, the F-measure criterion was introduced to reflect the harmonic mean of precision and recall. Therefore, the F-measure is a more important criterion than precision and recall because it includes both.

$$\text{Accuracy}= (TP+TN)/(TP+TN+FP+FN) \quad (2)$$

$$\text{Precision}= TP/(TP+FP) \quad (3)$$

$$\text{Recall}= TP/(TP+FN) \quad (4)$$

$$\text{F-meaure}=(2\times\text{Precision}\times\text{Recall})/(\text{Precision}+\text{Recall}) \quad (5)$$

4. 2. Hyperparameter Setting

Each of the classification algorithms, enumerated in section 3.5,

TABLE 2. Confusion matrix for gender identification

		Actual gender	
		Male	**Female**
Predicted gender	**Male**	True Positive (TP)	False Positive (FP)
	Female	False Negative (FN)	True Negative (TN)

requires one or more hyperparameters. Changing the hyperparameters greatly impacts the performance of the models; therefore, they must be adjusted carefully. In this research, we employed grid search for optimizing the hyperparameters. The optimal values of the hyperparameters obtained by the grid search are given in Table 3.

Depending on the features collected from users, various scenarios were conducted for evaluating the proposed method. In the following section, we use a unified notation for referring to different feature sets. Symbols C and D denote cumulative and detailed internet-based features, respectively, and symbol O indicates the other features (i.e., non-internet based features). Also, the asterisk at the top of these notations means that the feature selection step (i.e., random forest) is also performed. The remaining section describes scenarios one by one and provides the gender identification result, accordingly.

4. 3. Various Feature Sets Utilization Scenarios

- **Using cumulative Internet-based features**

In this scenario, different machine learning algorithms are applied and evaluated on the cumulative internet-based features. Table 4 shows the evaluation results of these classifiers.

As indicated in the table, these methods have not performed very well. This may be due to the lack of detailed internet features. The four applied features just show how the internet-based is used during the overall day, neglecting the details of using it within the day. The best performance of Table 4 is 0.53 in terms of precision. This prediction is not much different from the random guess. SVM with the radial basis kernel function has relatively high accuracy in comparison to others, but the other criteria of this method are not so high. On the other hand, the KNN outperformed other classifiers in terms of all criteria, which may be due to the simplicity of this classifier.

TABLE 3. The hyperparameter setting

Classifier	Hyperparameters value description
SVM	Regularization parameter = 0.025
SVM (Radial basis kernel)	Kernel variance = 10^{-7} Regularization parameter= 1
KNN	K=3
Decision Tree	Splitting criterion = gini-index
Random Forrest	Number of trees = 100, Splitting criterion = Gini-index, Number of features to consider for the best split = Sqrt of the number of features
AdaBoost	Number of weak classifiers = 100, Base classifier = Decision Tree
Naïve Bayes	Kernel = Gaussian

TABLE 4. Performance of gender identification on cumulative Internet-based features (FSC)

Algorithm	Accuracy	Recall	Precision	F-measure
LSVM	0.41	0.35	0.38	0.38
SVM (RBF)	**0.54**	0.4	0.4	0.4
KNN	**0.54**	**0.45**	**0.53**	**0.53**
Naïve Bayes	0.48	0.37	0.43	0.43
Decision Tree	0.48	0.38	0.43	0.43
Random forest	0.5	0.4	0.41	0.41
Adaboost	0.44	0.39	0.41	0.41

- **Using detailed Internet-based features**

In this scenario, classification algorithms are applied and evaluated on detailed internet-based features. The results are reported in Table 5. This table shows that among the various machine learning methods, Naïve Bayes has superior performance in predicting the gender of individuals according to 12 detailed internet-based features. After that, the decision tree obtaied the second-best place. In addition, by comparing Table 5 with Table 4, it can be concluded that the use of detailed internet-based features rather than cumulative ones provided more accurate information about the gender of individuals. This may be due to the use of more features providing more accurate information in the detailed internet-based features.

- **Using cumulative and detailed Internet-based features**

This scenario studies the effect of utilizing both detailed and cumulative internet-based feature sets in modeling. Table 6 shows the results of gender identification under this scenario.

According to this table, the performance of Naïve Bayes and decision tree have been better than other methods, in terms of accuracy. This result is consistent with the result of the second scenario reported in Table 5. Also, comparing Tables 5 and 6, indicates that the union of cumulative features and detailed ones has increased

TABLE 5. Performance of gender identification on detailed Internet-based features (FSD)

Algorithm	Accuracy	Recall	Precision	F-measure
LSVM	0.57	0.56	0.52	0.54
SVM (RBF)	0.52	0.3	0.2	0.24
KNN	0.55	0.48	0.6	0.53
Naïve Bayes	**0.72**	**0.68**	**0.65**	**0.66**
Decision Tree	0.62	0.67	0.5	0.57
Random forest	0.55	0.53	0.48	0.51
Adaboost	0.57	0.56	0.55	0.55

TABLE 6. Performance of gender identification on both cumulative and detailed Internet-based features (FSC+D)

Algorithm	Accuracy	Recall	Precision	F-measure
LSVM	0.57	0.56	0.52	0.54
SVM(RBF)	0.55	0.22	0.33	0.17
KNN	0.63	0.59	0.58	0.59
Naïve Bayes	**0.73**	**0.73**	**0.74**	**0.73**
Decision Tree	0.65	0.67	0.5	0.57
Random forest	0.6	0.57	0.55	0.56
Adaboost	0.57	0.55	0.57	0.56

TABLE 8. Performance of gender identification on all features (FSC+D+O)

Algorithm	Accuracy	Recall	Precision	F-measure
LSVM	**0.83**	**0.88**	**0.78**	**0.83**
SVM(RBF)	0.53	0.41	0.42	0.41
KNN	0.68	0.68	0.68	0.67
Naïve Bayes	0.73	0.73	0.7	0.72
Decision Tree	0.7	0.69	0.7	0.7
Random forest	0.71	0.6	0.57	0.58
Adaboost	0.65	0.63	0.63	0.63

the predictability and, consequently, the accuracy of the classifiers.

- **Using non-Internet-based features**

This scenario was conducted so that only non-internet-based features contribute to modelling the gender. Table 7 shows the results of this evaluation. It illustrates that Naïve Bayes has the highest accuracy among the classification algorithms, followed by random forest, LSVM, and Adaboodt, respectively.

- **Using All features**

In this scenario, all features, including cumulative internet-based, detailed internet-based, and non-internet-based features involved in the training phase. Table 8 reveals the performance evaluation of different classifiers under this scenario.

As it is inferred from Table 8, LSVM performed better than the other methods for all criteria, and Naïve Bayes is in second place. Furthermore, comparing Tables 7 and 8, it is apparent that the accuracy of all the classification algorithms has been improved by considering the internet features. This improvement determines the effect of using internet-based features along with other ones.

- **Summary of evaluating different feature sets**

To summarize the impact of applying different feature sets on the performance of gender identification models, we compared the best accuracy obtained for each of the enumerated scenarios. Figure 2 provides the result of this comparison. The chart shows that the detailed internet-based feature set gives better results than the cumulative one since it provides more accurate information. The use of all internet-based features has a better result than the use of detailed or cumulative internet features alone. Finally, the highest accuracy was achieved by utilizing all of the internet and non-internet-based features.

TABLE 7. Performance of gender identification on non-Internet-based features (FSO)

Algorithm	Accuracy	Recall	Precision	F-measure
LSVM	0.63	0.63	0.63	0.63
SVM(RBF)	0.53	0.53	0.26	0.35
KNN	0.51	0.51	0.49	0.47
Naïve Bayes	**0.70**	**0.70**	**0.70**	**0.70**
Decision Tree	0.55	0.55	0.54	0.54
Random forest	0.65	0.65	0.64	0.65
Adaboost	0.63	0.63	0.63	0.63

4. 4. Various Feature Sets Utilization Scenarios Along with Feature Selection

Applying the feature selection methods to the dataset before data modeling may improve the model's performance. Several experiments were conducted for evaluating the impact of feature selection on the model’s performance over different feature set utilizations. Experimental results are reported in subsequent sections.

- **Using Internet-based features with feature selection**

This scenario applies feature selection on the set of internet-based features (both cumulative and detailed) to select the impressive features for modeling the users gender. After performing the feature selection step, C1, C2, C3, C4, D3, D4, D5, D7, D10, and D1 were selected as the most important features. This indicates that all the

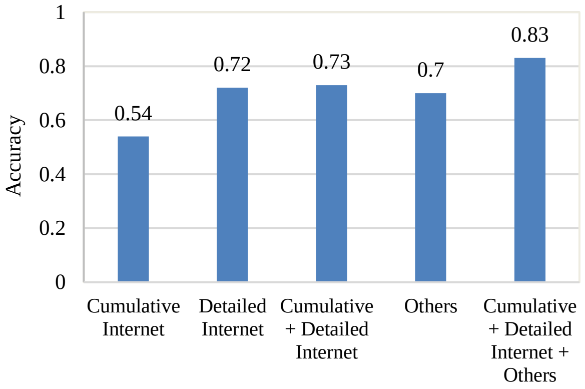

Figure 2. Summary of feature sets utilization

features of the cumulative set, including the average number of times the network data and Wi-Fi are turned on per day the average number of times Wi-Fi is turned on per day, the average number of network connections per day, the average duration of network connection per day, along with some detailed internet-based features including the average hours of internet use between 24 to 2 o'clock, 4 to 10 o'clock, 12 to 2 o'clock and 6 to 8 o'clock are important features for gender identification. Table 9 shows the evaluation results of the classification algorithms for the selected internet-based features.

Again, similar to the previous two scenarios, the Naïve Bayes method has superior performance by obtaining the accuracy of 0.70 in this scenario. Comparing Tables 9 and 6 indicates that the performance of Naïve Bayes has improved by applying the feature selection step. Moreover, the performance of Adaboost and decision tree have been slightly reduced. Random forest and KNN did not perform much differently. However, the performance of LSVM, SVM, and Naïve Bayes improved after applying the feature selection phase.

- **Using all features with feature selection**

This scenario includes feature selection applied to the whole set of features, including internet-based and non-Internet ones, prior to modeling. The feature selection algorithms suggested some non-internet-based features as well as D10 and D5 from the internet-based set. This selection reveals the duration of users' connection between 24 and 2 o'clock, and between 8 to 10 o'clock is affected by the user gender more than other features. The best obtained result is for LSVM in terms of accuracy, precision, recall, and F-measure. Table 10 shows the results of this evaluation. As inferred by the table, LSVM performed better than other classifiers and obtained the accuracy of 0.85. After that, random forest provided the best results with the 0.83 accuracy.

- **Summary of applying feature selection**

To analyze the effect of feature selection on gender identification accuracy, we compared the best accuracy obtained for each scenario. Figure 3 illustrates this comparison through a bar chart. As shown in this figure, applying feature selection on the Internet-based feature set has not resulted in significant improvement. But when all the features are used, applying feature selection has increased the accuracy by 0.02. Since the total number of features is large in this scenario and, irrelevant features may also exist, especially in the non-internet set, feature selection has positively affected the performance.

TABLE 9. Performance of gender identification on Internet-based features selected by the feature selection method (FS*C+D)

Algorithm	Accuracy	Recall	Precision	F-measure
LSVM	0.62	0.59	0.58	0.59
SVM(RBF)	0.55	0.48	0.28	0.36
KNN	0.62	0.63	0.53	0.58
Naïve Bayes	**0.7**	**0.71**	**0.65**	**0.68**
Decision Tree	0.57	0.55	0.5	0.52
Random forest	0.59	0.53	0.57	0.55
Adaboost	0.55	0.62	0.45	0.52

TABLE 10. Performance of gender identification on all features selected by the feature selection method (FS*C+D+O)

Algorithm	Accuracy	Recall	Precision	F-measure
LSVM	**0.85**	**0.85**	**0.85**	**0.85**
SVM(RBF)	0.64	0.57	0.68	0.62
KNN	0.7	0.7	0.71	0.69
Naïve Bayes	0.79	0.74	0.77	0.75
Decision Tree	0.66	0.55	0.67	0.56
Random forest	0.83	0.83	0.84	0.83
Adaboost	0.76	0.77	0.76	0.76

4. 5. Summary of Various Scenarios of the Proposed Method

To sum up, by comparing the four criteria of accuracy, recall, precision, and F-measure calculated for different machine learning methods in different scenarios, it can be concluded that LSVM has the best performance on all feature sets. These comparisons are presented in Figures 4 to 7 for different criteria and scenarios.

It can be concluded that if only the cumulative internet-based features are used, KNN is the best machine learning method for gender identification. But, if the detailed Internet-based features are provided, Naïve Bayes has the best performance. Moreover, if the non-internet-based features are considered in addition to

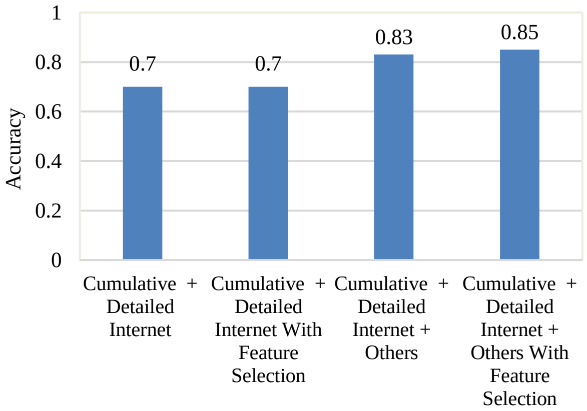

Figure 3. The effect of feature selection on gender identification accuracy for different feature sets

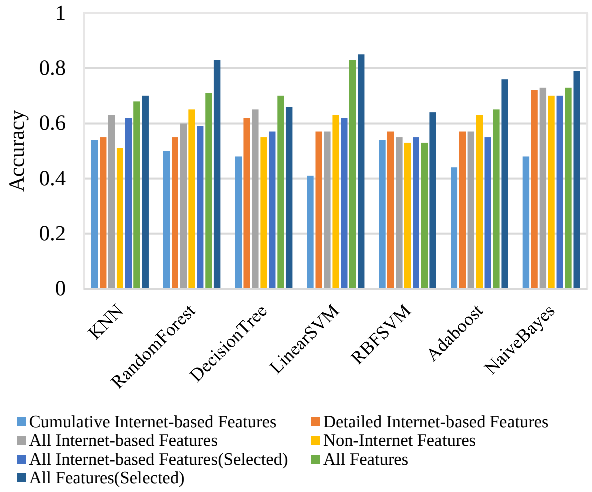

Figure 4. Accuracy of different classifiers in different scenarios

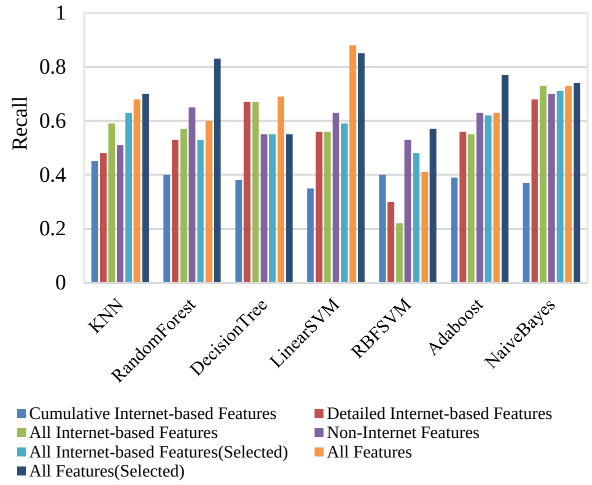

Figure 5. Recall of different classifiers in different scenarios

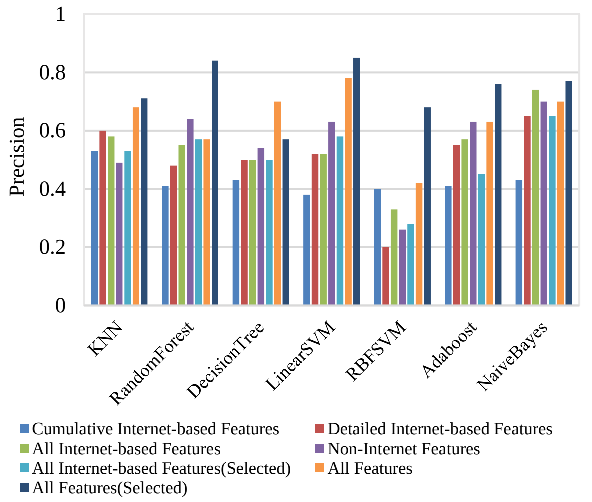

Figure 6. Precision of different classifiers in different scenarios

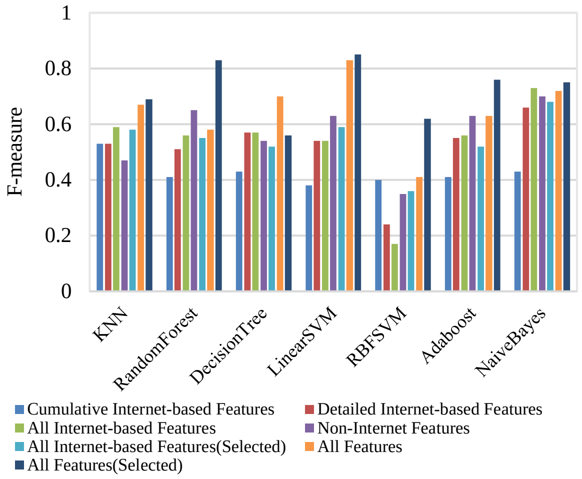

Figure 7. F-measure of different classifiers in different scenarios

internet-based ones, LSVM can better identify the individuals' gender. In addition, it can be inferred from the figures that applying the feature selection step improves the classification performance in most cases. The presence of internet-based features among the selected features shows that internet-based features have played a significant role in identifying the gender of individuals.

Finally, by comparing the present study with previous studies that used the same classifiers on a set of other features, it can be concluded that our utialized features have improved the performance of machine learning methods in gender identification. Therefore, the use of internet-based features along with other behavioural information of mobile phone users can lead to more accurate gender identification.

4. 6. Summary of Various Machine Learning Techniques

This section aims to compare the ability of various machine learning techniques in classifying genders over different feature selection scenarios. The average performance of each classifier over different scenarios is depicted in Figures 8 to 11, in terms of four performance criteria of accuracy, recall, precision, and F-measure. From these figures, it can be concluded that Naïve Bayes has the best and RBFSVM has the worst performances in gender detection for all criteria. The superiority of the Naïve Bayes method may be due to the simplicity of this classifier which leads to more regularization and generalization ability.

4. 7. Comparison with Related Works on Social Networks

To better evaluate the performance of the proposed method, it was compared with some important previous studies. To our best knowledge, no study has utilized internet usage data for gender identification. Therefore, the performance of the

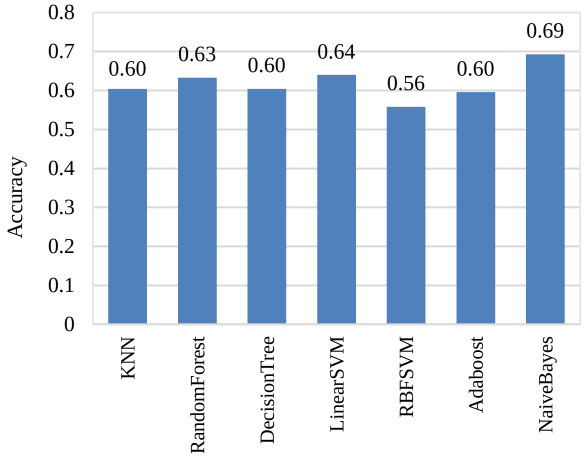

Figure 8. The overall accuracy of different machine learning techniques

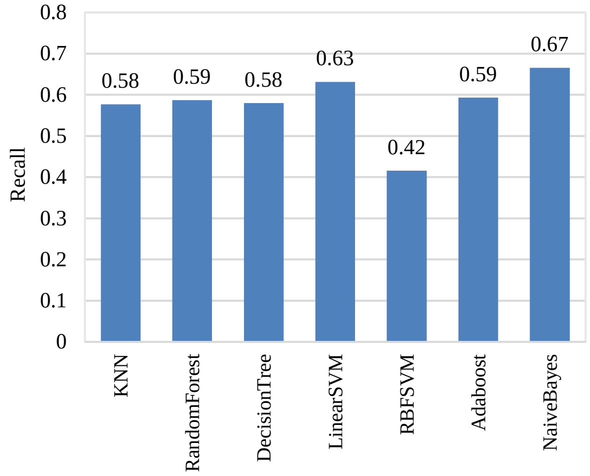

Figure 9. The overall recall of different machine learning techniques

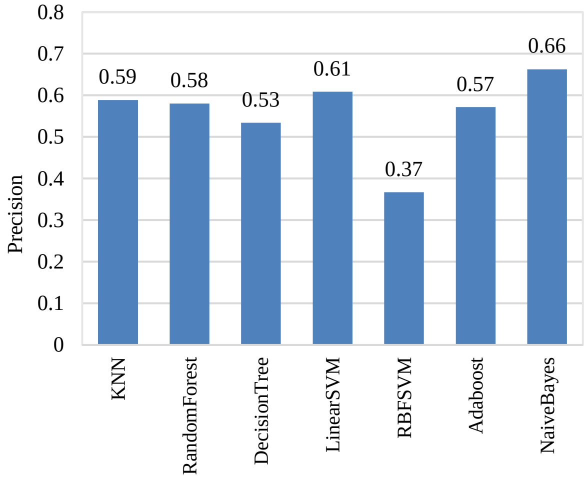

Figure 10. The overall precision of different machine learning techniques

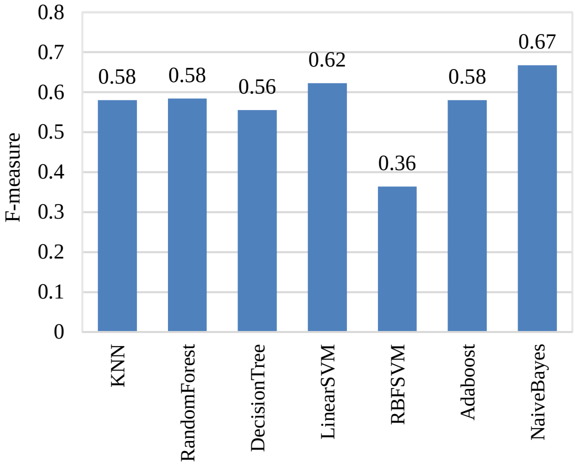

Figure 11. The overall F-measure of different machine learning techniques

proposed method is compared with the works that used other features for this task. This section compares the proposed method to gender identification studies in social networks. We choose machine learning classifiers the same as those used in the related studies to facilitate the comparison. But the dataset used in the proposed method is different from the datasets used in the previous studies. Related work contributed to the comparison includes Vashisth and Meehan [14] that determined the gender of individuals from individuals' tweets. Abadin et al. [15], which identified the gender of social network users using their posts on social networks. Table 11 reveals the results of the comparison. Since this section aims to compare with other studies, only the results related to the classifiers used in one of these two studies have been considered. Also, because previous studies have reported only the accuracy criterion, the results are reported in terms of accuracy in the table.

As mentioned earlier, in the proposed method, the best performance was achieved by applying the LSVM to all internet and non-internet-based features along with the use of feature selection. As shown in Table 11, the accuracy of this method is far better than the accuracies of the related works reported in literature [14, 15]. The same is true for Naïve Bayes, random forest and Adaboost classifiers. As a result, it can be inferred that the use of internet and non-internet data collected in this study has increased the efficiency of machine learning methods in identifying users gender.

4. 8. Comparison with Related Work on Mobile Phone Data To have a fair and comprehensive comparisons, in addition to studies that use social media information to identify the gender, the results of the proposed method is compared with other methods that use mobile phone information. Important studies from recent years have been selected to compare with the

TABLE 11. Comparing the accuracy of the proposed method with the related work in social networks

Method/Classifier	LSVM	Random Forest	Adaboost	Naïve Bayes
Proposed method (FS_{C+D})	0.57	0.6	0.57	0.73
Proposed method (FS^*_{C+D})	0.62	0.59	0.55	0.7
Proposed method (FS_{C+D+O})	0.83	0.71	0.65	0.73
Proposed method (FS^*_{C+D+O})	0.85	**0.83**	**0.76**	**0.79**
Abadin et al.[15] (all features)	-	0.6	0.55	-
Abadin et al.[15] (all features except text topic)	-	0.66	0.65	-
Abadin et al.[15] (less important features)	-	0.67	0.65	-
Abadin et al.[15] (selected features)	-	0.65	0.65	-
Vashisth et al. [14] (TF-IDF)	0.52	0.47	0.55	0.53
Vashisth et al. [14] (W2Vec)	0.52	0.47	0.55	-
Vashisth et al. [14] (GloVe)	0.52	0.48	0.52	-

TABLE 12. Comparing the accuracy of the proposed method with the related work on mobile phone data

Method	Accuracy
Proposed method(FS^*_{C+D+O})	0.85
Jain and Kanhangad [4]	**0.90**
Jain and Kanhangad [3]	0.82
Miguel-Hurtado et al.[10]	0.69

proposed method. Jain and Kanhangad [4] used collected signals from smartphone accelerometer and gyroscope sensors to predict gender. This method was evaluated in different scenarios based on the walking speed, the underlying device, and the sensors type. The accuracy for fast, normal, and slow walking as well as the two devices S-II and Note-II, in the case of using all sensors are averaged and reported in Table 12.

In another study [3], the same authors used touch screen gesture information along with various machine learning methods. They showed that the KNN obtains the best performance in this regard. Also, some researchers [10] used the information of sweep gestures to identify the gender. The accuracy of the best presented model has been considered in Table 12.

It is evident that the proposed method using all the features and feature selection could reach a higher accuracy than the work of Jain and Kanhangad [3] with the KNN model and the method of Miguel-Hurtado et al. [10]. But the best performance was achieved in another work of Jain and Kanhangad [4] among the compared studies. However, their presented model requires the use of accelerometer and gyroscope sensors and walking activity which threats the generality and universality of the deployment of the method. For example, the method does not apply to users who have been idle for a long time or on sensorless mobile phones. They are also device-dependent, and their generalizability to other devices with different sensors can be quite challenging. On the other hand, in their proposed method a high volume of data is recorded from gyroscope and accelerometer sensors that are not easy to store and process for the mobile phone as a small processing unit.

4. 9. Analysis of Gender-related Characteristics on Internet Usage

The evaluation results of the proposed method confirm that it is possible to identify the gender of users based on their internet usage. This result is in line with psychological studies conducted in this regard.

The gender-related characteristics which distinguish the internet usage pattern in females and males have been investigated in former studies [12]. Some of these distinguishing features that have been mentioned in psychological studies are as follows:

- The amount of family supervision is often more for female teenagers than males, which prevents them from spending too much time on the internet [28].
- Internet availability is higher for males than for females approximately all over the world [29].
- According to FMRI image analysis [30], males' and females' brains have different susceptibilities to internet gaming addiction.
- Sociocultural customs and norms cause different types of behaviour in females and males for example by imposing more restrictions on females for using the internet [31]. Again, Becker et al. [32] claimed that the effect of social and legal constraints on females is usually greater than on males.
- It was also stated [33] that excessive internet usage in males can be interpreted as an escape into cyberspace. It is a kind of self-medication behaviour for them in reaction to depression.

5. CONCLUSION

In this study, gender identification according to internet usage patterns was investigated. To this aim, two sets of Internet-based features (i.e., cumulative internet-based features and detailed internet-based features) were introduced to be used beside the non-internet-based features. Then, several models were obtained to predict the gender of mobile phone users based on internet and non-internet based feature sets. Various experiments were conducted to investigate the effect of cumulative and detailed internet-based features on the performance of gender identification models. From the experiments, it was inferred that using internet-based features along with other interaction-based ones can improve the performance of the models. The results of applying machine learning algorithms on the collected phone interaction data suggested that LSVM with the accuracy of 85% obtained the best results in identifying the gender of users. One of the limitations of this work is related to mobile phone computing power. Because of this limitation, it may not be possible to process the features on some types of smartphones and the features need to be sent to the server to be processed. Another considerable limitation of the presented work is that the collected dataset is limited in terms of both records and features. For future studies, more diverse feature sets (e.g., the data volume transferred in each connection or period, applications used by the users, etc.) can be introduced to represent the internet usage pattern of users. Other feature selection methods and machine learning algorithms can also be examined on the collected data to investigate if the results improve. In addition, the data extracted from social media can be combined with the smartphone usage data to improve the accuracy of the models. In addition to gender, the usefulness of internet usage features to identify other user characteristics such as the age and education level can be examined.

6. REFERENCES

1. Daneshvar, S. and Inkpen, D., "Gender identification in twitter using n-grams and lsa.", in Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), (2018).
2. Giannakopoulos, O., Kalatzis, N., Roussaki, I. and Papavassiliou, S., "Gender recognition based on social networks for multimedia production." in 2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop, IVMSP 2018 - Proceedings. Institute of Electrical and Electronics Engineers Inc, (2018).
3. Jain, A. and Kanhangad, V., "Gender recognition in smartphones using touchscreen gestures." ***Pattern Recognition Letters***, Vol. 125, (2019), 604-611. https://doi.org/10.1016/j.patrec.2019.06.008
4. Jain, A. and Kanhangad, V,, "Gender classification in smartphones using gait information." ***Expert Systems & Application*s**, Vol. 93, (2018), 257-266. https://doi.org/10.1016/j.eswa.2017.10.017
5. Meena, T. and Sarawadekar, K., "Gender recognition using in-built inertial sensors of smartphone.", in IEEE Region 10 Annual International Conference, Proceedings/TENCON. Institute of Electrical and Electronics Engineers Inc., (2020), 462-467.
6. Nguyen-Quoc, H. and Hoang, VT., "Gender recognition based on ear images: A comparative experimental study.", in 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2020. Institute of Electrical and Electronics Engineers Inc., (2020), 451-456.
7. Jain, A. and Kanhangad, V., "Investigating gender recognition in smartphones using accelerometer and gyroscope sensor readings", in 2016 international conference on computational techniques in information and communication technologies (ICCTICT). IEEE, (2016), 597-602.
8. Sarraute, C., Blanc, P. and Burroni, J., "A study of age and gender seen through mobile phone usage patterns in mexico.", in 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014). IEEE, (2014), 836-843.
9. Choi, Y., Kim, Y., Kim, S., Park, K. and Park, J., "An on-device gender prediction method for mobile users using representative wordsets". ***Expert Systems & Applications***, Vol. 64, (2016), 423-433. https://doi.org/10.1016/j.eswa.2016.08.002
10. Miguel-Hurtado, O., Stevenage, S. V., Bevan, C. and Guest. R. "Predicting sex as a soft-biometrics from device interaction swipe gestures." ***Pattern Recognition Letters***, Vol. 79, (2016), 44-51
11. Ramirez-Correa, P.E., Rondan-Cataluña, F.J. and Arenas-Gaitán, J. "Predicting behavioral intention of mobile Internet usage." ***Telemat Informatics***, Vol. 32, (2015), 834-841. https://doi.org/10.1016/j.tele.2015.04.006
12. Su, W., Han, X., Jin, C., Yan, Y. and Potenza, M.N., "Are males more likely to be addicted to the internet than females? A meta-analysis involving 34 global jurisdictions." ***Computers in Human Behavior***, Vol. 99, (2019), 86-100. https://doi.org/10.1016/j.chb.2019.04.021
13. Okazaki, S. and Hirose, M., "Does gender affect media choice in travel information search? On the use of mobile Internet.", ***Tourism Management***, Vol 30, (2009), 794-804. https://doi.org/10.1016/j.tourman.2008.12.012
14. Vashisth, P. and Meehan, K., "Gender classification using Twitter text data.", in 2020 31st Irish Signals and Systems Conference (ISSC). IEEE, (2020), 1-6.
15. Piot-Perez-Abadin, P., Martín-Rodilla, P. and Parapar, J., "Experimental analysis of the relevance of features and effects on gender classification models for social media author profiling." in ENASE., (2021), 103-113.
16. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., "Lightgbm: A highly efficient gradient boosting decision tree" ***Advances in Neural Information Processing Systems***, Vol. 30, (2017).
17. Kowsari, K., Heidarysafa, M., Odukoya, T., Potter, P., Barnes, L.E. and Brown, D.E., "Gender detection on social networks using ensemble deep learning.", in Proceedings of the Future Technologies Conference. Springer, (2020), 346-358.
18. Safara, F., Mohammed, A.S., Potrus, M.Y., Ali, S., Tho, Q.T., Souri, A., Janenia, F. and Hosseinzadeh, M., "An author gender detection method using whale optimization algorithm and artificial neural network.", ***IEEE Access***, Vol. 8, (2020), 48428-48437
19. Sadeghian, A. and Kaedi, M.. "Happiness recognition from smartphone usage data considering users' estimated personality traits.", ***Pervasive and Mobile Computing***, Vol. 73, (2021), 101389
20. James, G., Witten, D., Hastie, T. and Tibshirani, R., An introduction to statistical learning, Springer, (2013).
21. Breiman, L., Classification and regression trees, Wadsworth Stat Ser 358, 1984.
22. Han, J., Pei, J. and Kamber, M., Data mining: concepts and techniques, Elsevier, 2011.
23. Jo, T., Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning, Springer Nature, (2021).
24. Yadav, S. and Shukla, S., "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality

classification.", in 2016 IEEE 6th International conference on advanced computing (IACC). IEEE, (2016), 78-83.

25. Hamidi, H. and Rafebakhsh, M. S. “Analyzing factors influencing mobile social media marketing acceptance among customers.”, ***International Journal of Engineering, Transactions C: Aspects***, Vol. 35, No. 6, (2022), 1209-1216. DOI: 10.5829/IJE.2022.35.06C.13

26. Yavari, A., Hassanpour, H., Rahimpour Cami, B. and Mahdavi, M. “Election Prediction Based on Sentiment Analysis using Twitter Data.” ***International Journal of Engineering, Transactions B: Applications***, Vol. 35, No. 2, (2022), 372-379. DOI: 10.5829/IJE.2022.35.02B.13

27. Jaderyan, M. and Khotanlou, H. “Automatic hashtag recommendation in social networking and microblogging platforms using a knowledge-intensive content-based approach.”, ***International Journal of Engineering, Transactions B: Applications***, Vol. 32, No. 8, (2019), 1101-1116. doi: 10.5829/ije.2019.32.08b.06

28. Yu, L. and Shek, D. T. L.. “Internet addiction in Hong Kong adolescents: A three-year longitudinal study.” ***Journal of Pediatric and Adolescent Gynecology***, Vol. 26, No. 3, (2013), 10-17. https://doi.org/10.1016/j.jpag.2013.03.010.

29. International Telecommunication Union “ICT facts and figures 2016”. (2016), from https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016.pdf.

30. Wang, Z., Hu, Y., Zheng, H., Yuan, K., Du, X. and Dong, G. “Females are more vulnerable to Internet gaming disorder than males: Evidence from cortical thickness abnormalities.” ***Psychiatry Research: Neuroimaging***, Vol. 283, (2019), 145-153. https://doi.org/10 .1016/j.pscychresns.2018.11.001.

31. Hafkin, N. J. and Huyer, S. “Women and gender in ICT statistics and indicators for development.” ***Information Technologies and International Development***, Vol. 4, No. 2, (2007), 25-41.

32. Becker, J. B., McClellan, M. and Reed, B. G. “Sociocultural context for sex differences in addiction.” ***Addiction Biology***, Vol. 21 No. 5, (2016), 1052-1059. https://doi.org/10.1111/adb.12383.

33. Jang, M. H. and Ji, E. S. “Gender differences in associations between parental problem drinking and early adolescents' Internet addiction.” ***Journal for Specialists in Pediatric Nursing***, Vol. 17, No. 4, (2012), 288-300. https://doi.org/10.1111/j.1744-6155.2012.00344.x

Persian Abstract

چکیده

جنسیت افراد بخش مهمی از هویت افراد را تشکیل می‌دهد. تشخیص جنسیت کاربران سیستم‌ها، برای شخصی‌سازی خدمات و پیشنهادات در بسیاری از کاربردها مفید واقع می‌گردد. از طرف دیگر، امروزه اغلب افراد ساعات زیادی را با تلفن همراه خود سپری می‌کنند. پژوهش‌ها نشان داده است که نحوه تعامل کاربران با تلفن همراه، تحت تاثیر جنسیت آنها است. اما سیستم‌هایی که تا کنون برای تشخیص جنسیت کاربران بر اساس تعاملات آنها با تلفن همراه ارائه شده‌اند، یا دقت کافی در تشخیص ندارند و یا نیاز به سنسورها و فعالیت خاص کاربر برای تشخیص جنیست دارند و بنابرین قابلیت استفاده در حالت کلی و عمومی را ندارند. در این پژوهش، برای اولین بار به تشخیص جنسیت افراد بر اساس تعاملات آن‌ها با تلفن همراه، و به‌طور ویژه بر اساس ویژگی‌های مربوط به الگوهای استفاده آنها از اینترنت پرداخته شده است. بدین منظور با در نظر گرفتن یک نمونه تصادفی از افراد جامعه، اطلاعات دموگرافیک، الگوهای استفاده از اینترنت و همچنین ویژگی‌های مربوط به تعامل کاربران با تلفن همراه به صورت خودکار توسط یک برنامه کاربردی ثبت شده است. سپس جنسیت کاربران بر اساس ویژگی‌های استخراج شده و با به کارگیری روش‌های مختلف یادگیری ماشین مدل‌سازی شده است. نتایج ارزیابی‌ها نشان داد که ویژگی‌های اینترنتی نقش مهمی در شناسایی جنست کاربران تلفن های همراه دارند و روش ماشین بردار پشتیبان خطی با اجرا بر روی تمام ویژگی‌ها با اعمال انتخاب ویژگی با صحت 85% و شاخص-اف ٪۸۵ بالاترین عملکرد را در شناسایی جنسیت کاربران داشته است.