Gender Inference for Facebook Picture Owners

Bizhan Alipour, Abdessamad Imine, Michaël Rusinowitch

Gender Inference for Facebook Picture Owners*

Bizhan Alipour, Abdessamad Imine and Michaël Rusinowitch

[1] Lorraine University, Cnrs, Inria, 54506 Vandœuvre-lès-Nancy, France
firstname.lastname@loria.fr

Abstract. Social media such as Facebook provides a new way to connect, interact and learn. Facebook allows users to share photos and express their feelings by using comments. However, Facebook users are vulnerable to attribute inference attacks where an attacker intends to guess private attributes (e.g., gender, age, political view) of target users through their online profiles and/or their vicinity (e.g., what their friends reveal). Given user-generated pictures on Facebook, we explore in this paper how to launch gender inference attacks on their owners from pictures meta-data composed of: (i) alt-texts generated by Facebook to describe the content of pictures, and (ii) comments posted by friends, friends of friends or regular users. We assume these two meta-data are the only available information to the attacker. Evaluation results demonstrate that our attack technique can infer the gender with an accuracy of 84% by leveraging only alt-texts, 96% by using only comments, and 98% by combining alt-texts and comments. We compute a set of sensitive words that enable attackers to perform effective gender inference attacks. We show the adversary prediction accuracy is decreased by hiding these sensitive words. To the best of our knowledge, this is the first inference attack on Facebook that exploits comments and alt-texts solely.

Keywords: social network, privacy, inference attack, gender inference, picture.

1 Introduction

In attribute inference attacks, the attacker aims to guess user's private attributes (such as gender, age, political view, or sexual orientation) from user's publicly available data. The attacker can train machine learning classifiers based on the collected data in order to infer private attributes. Recent works have investigated friend-based [9] and behavior-based [21] inference attacks on Facebook users.

Friend-based attacks are based on the intuition that friends share similar attributes. Indeed, the attacker proceeds in two steps: (i) drawing up the friend list of the target user; (ii) computing correlations from the public data of the target user and his/her friends in order to identify hidden attributes of the target user [8]. As for Facebook behavior-based attacks, they are based on the intuition that *you are how you behave*. Thus, the attacker examines user behavior from liked

* This work is supported by DIGITRUST (http://lue.univ-lorraine.fr/fr/article/digitrust/).

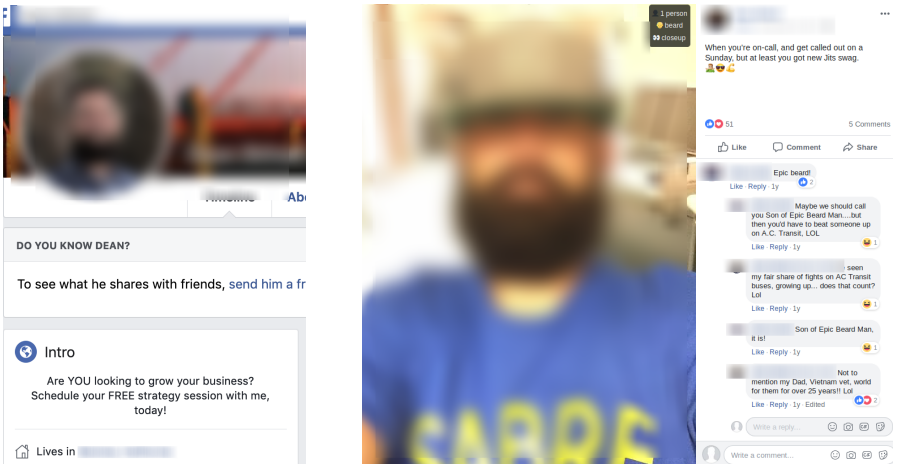
pages and joined groups in order to infer his/her private attributes [2]. However, in a real scenario, the amount of available information to an attacker is rather small. In this work, we focus on Facebook as it is the largest social network in the world. More precisely, we consider the gender inference problem as gender can be considered as a sensitive information. Indeed, users wish to hide their gender for a variety of reasons: 1) They wand to strengthen protection against discrimination. For instance, setting the gender to female results in getting fewer instances of an ad related to high paying jobs than setting it to male [5]. Targeting by sex is just one way Facebook and other tech companies let advertisers focus on certain users and exclude others. Facebook lets advertisers spend only on those they want to reach [17]; 2) They use the protection of anonymity to reduce the social risks of discussing unpopular and taboo topics [1,22]; 3) They want to prevent any form of sexual harassment and stalking as reported by the survey participants in [10].

Picture alt-text and comments. Unlike previous works, we attempt to learn the target user gender based on his/her online pictures. Note that publishing pictures enable their owners to increase connectivity and activity on social networks. Nevertheless, they lose privacy control on their pictures because of some information (*meta-data*) added during online publication, such as: (i) automatic alt-text (AAT) included by Facebook platform to describe the content of pictures, and (ii) comments posted by the closest friends as well as by strangers. Facebook launches automatic alt-text (AAT), a system that applied computer vision technology to identify faces, objects, and themes from pictures displayed by Facebook users to generate descriptive alt-text that can be used by the screen reader. They proposed the system to help blind people to feel more connected and involved in Facebook. The alt-text always starts by *Image may contain* and is followed by a list of recognized objects. Facebook provides a list of 97 objects and themes that provides different sets of information about the image, including people *(e.g., people count, smiling, child, baby)*, objects *(e.g., car, building, tree, cloud, food)*, settings *(e.g. inside restaurant, outdoor, nature)*, and themes *(e.g., close-up, selfie, drawing)* [20]. On the plus side, AAT provides free, additional information about photos, and makes blind people feel more included and engaged in photos. On the negative side, this technology also provides the social network with yet another entry point in user private life, namely pictures. Furthermore, when observing a picture on Facebook, people write instinctively comments to express their feeling about the picture. These comments contain potentially sensitive information and are often available to the attacker.

Motivation. In order to raise awareness of social network users about their privacy, we propose to show the possibility of gender inference attack based on seemingly innocent data: *alt-text* which is generated by Facebook and picture *comments* which are written by target friends, friends of friends or ordinary users. We determine how these pieces of information lead to design feature sets that can be processed by an attacker to infer gender of picture owners. Users

can hide the information on their profile Figure 1(a), and the pictures can be still available to public.

Unlike existing inference attacks that require a exhaustive search in the target user networks, groups, and pages that he/she liked, our attack only needs target user public pictures. Additionally, unlike twitter gender inference attack that is based on target user writing sample [13], we intend to launch gender inference attack even when the users writing style is unavailable. For instance in Figure 1(a), the user has hidden most of his/her attributes on the profile, and *location* is the only available attribute to the attacker. Figure 1(b) shows the possibility of gender inference attacks based on alt-text and comments. In this example, Facebook generates the alt-text by detecting *one person* in the picture which has a *beard* and takes the picture in a *close-up* theme. The detection of *one person* guarantees that the alt-text and comments linked to the picture are pointing to one person. Hence, the presence of *beard* in alt-text, *son of epic, beard man* in comments can lead to gender inference attack. This attack target users who concern about their privacy (users who hide any types of available information such as friend list, liked pages, groups, and attributes on the profile) but they do not consider pictures meta-data (which are coming from Facebook and friends, friends of friends and ordinary person) as a harmful information.



(a) (b)

Fig. 1: User profile: (a) Hiding attributes (b) alt-texts and comments in picture.

We notify two observations. First, the alt-text is useful to filter/select the most informative image for the gender inference attack. (Subsection 4.2). Second, sensitive information can be extracted from comments.

Outline. The paper is organized as follows: we review related work in Section 2. In Section 3 we describe our gender inference problem and overview our framework to analyze pictures. Section 4 presents in detail our methodology. Section 5 discusses our experimental results. We conclude the paper in Section 6.

2 Related Work

In this section, we review a number of recent studies that demonstrated attribute inference attacks on Facebook and Twitter.

Hidden attribute inference attacks on Facebook. In [2], the authors inferred users private attributes using the public attributes of other users sharing similar interests and an ontologized version of Wikipedia. The authors of [7] proposed a new privacy attacks to infer attributes (e.g., locations, occupations, and interests) of online social network users by integrating social friends and behavioral records. They showed by increasing the availability of target user online information their results have serious implications in Internet privacy. In [8] the authors extended the Social-Attribute Network(SAN) framework with several leading supervised and unsupervised link prediction algorithms for link and attribute inferences. In [19] the authors show that a recommender system can infer the gender of a user with high accuracy, based solely on the ratings provided by users, and a relatively small number of users who share their demographics. Focusing on gender, they designed techniques for effectively adding ratings to a users profile for obfuscating the users gender, while having a nonsignificant effect on the recommendations provided to that user. To sum up, these inference attacks are costly as they assume that the entire or part of social network information is available to the attacker. Our work for gender inference does not explore the target user network and it relies only on small information (i.e. alt-text and comments of target user published pictures).

Twitter gender prediction. In [12], the authors focused on the task of gender classification by using 60 textual meta-attributes, for the extraction of gender expression linguistic in tweets written in Portuguese. Therefore, they considered characters, syntax, words, structure, and morphology of short length, multigenre, content-free texts posted on Twitter to classify authors gender via three different machine-learning algorithms as well as evaluate the influence of the proposed meta-attributes in the process. The work in [15] consists in identifying the gender of users on Twitter using perceptron and Naive Bayes from tweet texts. These works try to infer gender of tweet authors by semantically or syntactically analyzing the tweets. Unlike Twitter-based works, we discard the target user comments as they are irrelevant. Indeed, the target user is often careful to not disclose gender. More precisely, we try to infer gender indirectly by using comments underneath his/her pictures. These comments may be posted by friends, friends of friends or strangers.

3 Model Overview

In this section, we give the problem description and then present a brief overview of our framework.

Problem Description In this study we consider three different scenarios according to availability of data: In the first scenario, we only consider alt-texts dataset as input data to infer the target user gender. This scenario happens when the pictures do not receive any comment, or when the comments have been hidden by their owners. In the second scenario, we have only access to comments dataset. We assume pictures are publicly visibility and Facebook is unable to generate alt-texts for them. In the last scenario, we use both extracted comments and alt-texts dataset. We assume picture comments are publicly available and alt-text has computed for that picture. We also introduce some working hypothesis:
1) We do not have access to the gender of commenters who wrote the comments underneath pictures.
2) We ignore the comments written by the target, since we assume the target is clever enough to hide gender information in his/her own texts.
3) We do not know whether the commenters are target user friends, friends of friends, or ordinary persons.
4) We do not perform any computer image processing on the target pictures.
5) We do not consider user profile name as an input to our attack process. Although some names only used for a specific gender, the cultural, and geographic origin of names is known to have a great effect on the reliability of gender inference methods [14]. Moreover, the Facebook user may use the shortened name as a chosen name. It is due to privacy concerns, and it is a popular tactic to be identifiable only to friends, but not so easily to a stranger. Let us now discuss the difficulty that arises when we try to solve the problem. In some situations, comments and alt-text seem to convey contradictory information. Below, we represent two examples were the comments alone orient gender inference towards one value but checking alt-text revealed that the conclusion is wrong.
Image 1:
Generated alt-text: 1 person, smiling, child and closeup
Comment: He looks so damn happy
Image 2:
Generated alt-text: dog, outdoor and nature
Comment: Who got u that handsome

The first image comment contains masculine pronoun HE, orienting the labeling towards male gender. The presence of *child* in alt-text suggests that the comment is pointing to a baby which is boy. In the second image, the comment orients also the labeling towards male as *handsome* is more often used for men. However checking the alt-text suggests that there is only a dog inside the picture. In order to avoid these misleading situations we pre-process the collected datasets and filter the pictures according to some rules defined in Subsection 4.2.

Our Framework. Figure 2 depicts the five components (detailed in Section 4) of our framework for gender inference:

1. Data crawling. To perform attribute inference attack, we need to collect pictures meta-data such as alt-texts generated by Facebook and comments generated by friends, or friends of friends or ordinary users.
2. Data pre-processing. It is a technique to prune the extracted data. We apply two pre-processing steps. In the first step, we filter the pictures based on their generated alt-texts, and in the second step we prune the comments.
3. Features extraction. Feature extraction is the process of converting the input data into a set of features.
4. Feature selection. It is the process of selecting a subset of features which contribute most to the output.
5. Construction of a classification model to classify the target user gender.

The architecture is represented in Figure 2. We describe each component in the following sections.

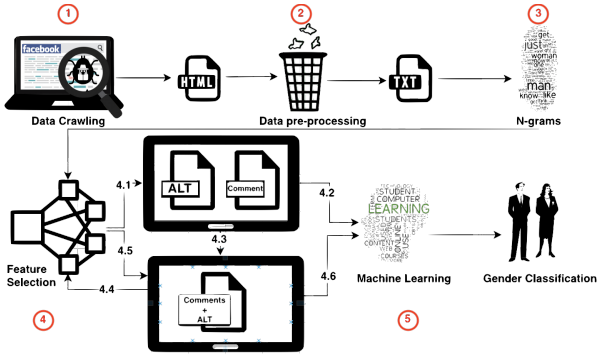


Fig. 2: Framework

4 Methodology

In this section, we describe in details our methodology for gender inference attack. Basically, we discuss the way of extracting the raw data from the internet, finding useful information from the extracted data and prune the extracted data in such a way that present better prediction accuracy.

4.1 Data crawling

Our objective is to infer the gender by collecting pictures meta-data. For each picture that we encounter, our crawler extracts the alt-text, and comments from the HTML file related to the picture. The gender of the user who posted the picture is collected (when available) , in order to create labeled datasets to be exploited by our supervised machine learning algorithms. Our data is labeled with two genders *female and male*, corresponding to biological sex. The data is derived from random Facebook users. Let U be the set of target user pictures $U = \{u_1, u_2, u_3, .., u_m\}$. For every picture we extract $u_i =< a_i, c_i >$ where a_i is an alt-text presented by Facebook and c_i is the set of comments for that

picture. We denote (i) $A = \bigcup_{0<i\leq m} a_i$ the set of all extracted alt-texts, and (ii) $C = \bigcup_{0<i\leq m} c_i$ the set of all extracted comments.

After constructing the comments and alt-texts dataset C and A respectively, we exploit these two datasets to launch gender inference attacks. It is worth mentioning that for each picture, either or both datasets may be empty. The possibility of an attack depends on how informative are alt-texts and comments.

4.2 Data pre-processing

Initially, we perform two pre-processing steps to clean collected data. We define a *singular form* to be a form that points only to one person. For example: *Handsome man* is a singular form that point to one man. A *plural form* is a form that pointing to more than one person. For example: *beautiful women* is a plural form. Now we discuss about pre-processing which consist of:

A: Filtering the picture w.r.t the following alt-texts rules Facebook provides various types of information about the image through alt-text, and they categorize this information into four tags including *people, objects, settings,* and *themes* (detailed in Section 1). In this works [3], for example, they showed emotional words such as *cute* are selected by their methods for images containing *dog* and *kid*. As we discussed in Section 3, it was problematic to select picture that contain useful meta-data. To solve the problem, we used *people* and *objects* tags inside picture generated alt-text.
Below, we describe the rules that we set to keep the pictures meta-data in our inference analysis.

Rule 0. As a default rule, we keep the picture meta-data if there is no *person* in *people tag*, no *animals* in *objects tag*, and no *child* inside the alt-text.

Rule 1. We keep the picture meta-data if the generated alt-text contains *1 person*. Note, there might be other tags such as *objects*, *setting*, and *themes*. This rule is satisfied when there is only *1 person* in *people tag* and no *animals* in *objects tag* no *child* in the generated alt-text.

Example 1. In the following example, we keep the picture meta-data as there is *1 person* in *people tag* and there is no other tags.
Generated alt-text: 1 person
Comments$_1$: Hot mom!
Comments$_2$: Thats a really good picture!

Rule 2. If the number of people mentioned by alt-text is more than *1* in *people tag*, then we analyze the comments without considering *animals* and *child* tags. We keep the picture meta-data if comments point only to one gender and contain at least one singular form. By having all male or all female words in the comments, we can be certain that people with the same gender are inside

the picture and we can assume that one of those people might be the owner. In another hand, If there are mixed gender in the picture, then we are not sure (for that picture) the owner is male or female.

Example 2. In the following example, according to alt-text, there are four people inside the picture. In this case, we consider the plural and singular form of words. Hence, the presence of a HANDSOME MAN which is pointing to one person leads us to keep the picture meta-data for further analysis. Note that in this example there is no word pointing to different gender (*female and male*).
Generated alt-text: 4 people, people smiling
$Comments_1$: Handsome man, And a yummy sunday
$Comments_2$: Wow! I never thought this day would happen

Rule 3. If alt-text contains *1 person* in *people tag* and some *animals* in *objects tag*, then we further analyze the comments. We keep the picture meta-data if comments orient to only one gender. An example of such a situation is given below:

Example 3. Generated alt-text: 1 person, dog and indoor
$Comments_1$: Sorry for your loss
Owner reply: Thank you
$Comments_2$: aw, i'm so sorry
$Comments_3$: RIP sweet Lady

B: Cleansing comments: In this work, we focus on English writing. So, in the second step we clean the comments as follow: (i) discarding non English comments by using python libraries such as *TextBlob, Guess_language*, and (ii) removing animated and graphic symbols such as *GIF*.

4.3 Features extraction

N-grams. Feature extraction is the process of converting the original data to a dataset that contains a reduced number of features. The extracted features comprise information related to the desired properties of the original data. Difficulty in analyzing data from social media raises from the presence of different kinds of textual errors, such as misspellings and grammatical errors. Traditional Natural Language Processing (NLP) techniques cannot always deal with texts that often do not follow even the simplest and most basic syntactic rules [16]. In order to capture syntactic similarities, we employ n-grams. Basically, n-grams are a set of co-occurring words within a given window n. The basic point of n-grams is that they capture the language structure from the statistical point of view, like which words are likely to appear together in *male/female comments/alt-text* dataset. Facebook users use the shortened form of a word (abbreviation) more often to save time on typing and expressing their feelings and thought in shortcuts inside comments. N-grams finds normal and deformed words that might be used more often by Facebook users. As a result, we received a feature *love u*, which contains a deformation letter u that refer to pronoun *you*. Note, the u in *love u* can be considered as a misspelled letter such as a by NLP techniques.

Optimum window size. Optimum n-grams length depends on the data type. For example, if the size of n in n-grams is too short, it may fail to capture important block of words. On the other hand, if it is too long, it may fail to capture the general knowledge [4]. For terminology extraction, Kenneth Church proposes a window size of 5 ($n \leq 5$). According to Church, this size is a good compromise: on one hand, it is large enough to show some semantic relationships between words, and on the other hand, it is not too large to lose the relationships that demand strict adjacency between words [6]. With that in mind, we employ n-grams on *alt-texts* and *comment* dataset separately to generate distinct feature set. Our experiments showed that introducing *5-grams* on *comments dataset* and *6-grams* on *alt-texts dataset* degraded the performance, so we kept the lengths up to *4-grams* and *5-grams* in *comments* and *alt-texts dataset* respectively. let F_c be the comments feature set containing (f, v) where f is the feature and v is the occurrence of the feature, then in this work, F_c is the sequence of n words, where $n \in \{1, 2, 3, 4\}$, generated by target user friends, friends of friends or ordinary users to express their feelings about the target user pictures. Respectively, Let F_a be the alt-texts feature set , then F_a is the sequence of n words, where $n \in \{1, 2, 3, 4, 5\}$, generated by Facebook. After extracting features from each dataset, we applied four feature selection algorithms to reduce the feature space and noise in the represented data.

4.4 Feature selection techniques

Feature Selection is the process of selecting features which contribute more to the prediction output (in our case, *male* and *female*). The choose of feature selection methods differ based on the problem and available data. Below, we discuss four feature selection methods that we have employed: *Chi-Square*[1] is used to test if the relationship of a dependent variable is significant to an independent variable. *Information Gain*[2] indicates the amount of information the independent variable presents with respect to the classification target attribute. It measures the difference in information was available before knowing the attribute value and after knowing the attribute value. *Feature importance*[3] provides a score for each feature, the higher the score more important or relevant is the feature towards the output variable.Feature importance is an inbuilt class that proceeds with Tree Based Classifiers. *Univariate feature selection*[4] examines each feature individually to determine the strength of the relationship of the feature with the response variable. Algorithm 1 describes our feature selection process. We introduce the following notations:

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html`

[2] `https://www.bogotobogo.com/python/scikit-learn/scikt_machine_learning_Decision_Tree_Learning_Informatioin_Gain_IG_Impurity_Entropy_Gini_Classification_Error.php`

[3] `https://www.scikit-yb.org/en/latest/api/features/importances.html`

[4] `https://scikit-learn.org/stable/auto_examples/feature_selection/plot_feature_selection.html`

1. LS is a document that associates to each occurrence of a feature the corresponding label (in our case *female, male*) of the profile from which this occurrence comes.
2. FSA_i where $i \in \{1, 2, 3, 4\}$ are feature selection algorithms used in this study.
3. $Score(i, f, LS)$ is the score of feature f with algorithm FSA_i.
4. T is a threshold. We keep the features that contribute more than T to the final result. We generally set the threshold to 0.5.
5. Res_i i $\in \{1, 2, 3, 4\}$ is the set of features with a score above the threshold.
6. MLA_m where $m \in \{1, 2, 3, 4, 5, 6\}$ are machine learning algorithms.
7. $Accuracy(m, F, L)$ measures the accuracy of MLA_m when applied to a set of features F and a labelling L of profiles.Accuracy is obtained as the ratio of correct predictions to the total of predictions.
8. $PS(J)$ is the intersection of feature sets obtained by selection algorithms with index in J.
9. $V(m, J)$ is the accuracy of machine learning algorithm MLA_m applied to PS_J and LS.
10. $(\bar{m}, \bar{J})$ is the combination of a feature set and a machine learning algorithm that gives the highest accuracy.

Note that FSA_i takes LS corresponding to F_c and F_a separately as input and generate score for each feature according to their predictive significance. Res_i contains features that are above the threshold. Using different feature selection methods lead to different selected features that might generate different prediction performance. In order to identify more representative features for better prediction, we evaluate all the possible individual and combined feature selection methods [18,11]. Later we run machine learning algorithms in order to find the accuracy of each individual and combined feature selection algorithms. Finally, we select the feature set and machine learning algorithm that give the highest accuracy. We used six machine learning algorithms such as *Logistic Regression*, *Random Forest*, *K-Nearest Neighbors*, *Support Vector Machine*, *Naive Bayes*, and *Decision Tree* to compare the performance of each individual and combined feature selection algorithms.

5 Experimental Result

We first describe the datasets that we used for the evaluation, followed by the representation of machine learning results in different scenarios.

5.1 Experimental Setup

Datasets Using a python crawler, we collected a set of 3,500 pictures. We note that among those pictures, Facebook was unable to generate alt-text for *200* pictures. Moreover, we collected *16,935* comments from *3,500* single pictures, among which *400* pictures did not receive any comment.

input : LS, T, $FSA_i, (1 \leq i \leq 4)$, $MLA_m, (1 \leq m \leq 6)$
output: Best Features Set BFS

Step1:
for *all* i **do**
  $Res_i \leftarrow \{f \mid score(i, f, LS) \geq T\}$
end
Step 2:
for *all* $J \subseteq \{1,2,3,4\}$ **do**
  $PS(J) \leftarrow \bigcap_{i \in J} Res_i$
  for *all* m **do**
    $V(m, J) \leftarrow Accuracy(m, PS(J), LS)$
  end
end
$(\bar{m}, \bar{J}) \leftarrow Argmax\ \{V(m, J)\}$
$BFS \leftarrow PS(\bar{J})$

Algorithm 1: Best feature selection algorithm

Evaluation Metrics We evaluate our attack using the standard *Accuracy*, *Precision*, *Recall*, and *F1_score* metrics. Below, we describe each one briefly.
Accuracy: The attackers output has two classes *Male* and *Female*. Accuracy is the fraction of the correct predictions (predicting male as male and female as female) for unknown data points.
Recall: It refers to the percentage of total relevant results that correctly classified. Recall is defined as the number of true positives divided by the number of true positives plus the number of false negative.
Precision: Precision means the percentage of results which are relevant. Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.
F1_score: It takes precision and recall into account to evaluate the model.

Experiments We considered the gender inference attack as a binary classification problem. To that end, we have applied well-known supervised classification algorithms in our work. To select the suitable classifier for our project, we tested several supervised machine learning algorithms such as *Logistic Regression*, *Random Forest*, *K-Nearest Neighbors*, *Support Vector Machine*, *Naive Bayes* and *Decision Tree*. Experiments carried out for *alt-texts feature set*, *comments feature set*, and *combined feature set*. In order to evaluate the classifier, we selected the same number of male and female to prevent biased classification. The evaluation was conducted by splitting the dataset into train/test size. Train-test splitting was preferable in this study as it runs k-times faster than k-fold. We vary the size of the training set to measure the difference in the attack accuracy. Finally, we choose the train-test size of *70-30* which gives the best accuracy. The following experimental result obtained by using classifiers implemented in Python library *scikit-learn*. Below, we present our results for each scenario.

5.2 Alt-texts feature set

Here, we discuss the scenario where we have only access to the alt-texts dataset. In this scenario, the input of Algorithm 1 is the corresponding *LS* to alt-texts feature set. Figure 3 displays our inference results on alt-texts dataset. In addition to accuracy, we also show the result of *Precision*, *Recall*, and *F1-Score*. According to Algorithm 1, the intersection of *Feature importance*, and *Univariate feature selection* perform the best and generate 68 features. The efficacy of the selected features was measured by creating a classifier which only used these features *(red bars)*. According to Figure 3, *Logistic Regression* performs slightly better than *Decision Tree* classifier in *Accuracy*.Based on the result, an attacker can infer the target user gender with accuracy of 84%. Finding important metrics depend entirely on the problem. As the target variable classes *(female, male)* in our dataset are balanced, then we can use *Accuracy* as an important metrics in our project. Note, *F1-Score* take the harmonic mean between *Precision* and *Recall*. So, we can observe that *Logistic Regression* performs better than other supervised classifiers in the alt-texts dataset Based on *Accuracy* and *F1-Score* sub-graphs. *Logistic Regression* also perform slightly worse than *Support Vector Machine* in *Recal* sub-graph. In the other hand, Figure 3 *(blue bars)* represent a significant deterioration in *Accuracy*, *Recal* and *F1-Score*

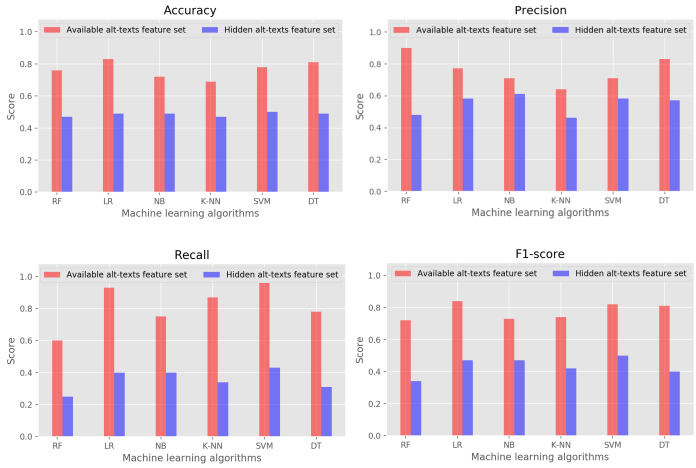


Fig. 3: Accuracy, Precision, Recall, and F-Score evaluation of alt-texts dataset

5.3 Comments feature sets

In this subsection, we demonstrate the interest of comments for performing inference attacks. In this scenario, we have only access to the comments dataset, and the input of the Algorithm 1 is the corresponding *LS* to comments feature set. According to Algorithm 1, the intersection of *Chi-Square*, *Feature importance*, and *Univariate feature selection* perform the best and create 66 features. Figure 4 displays our inference attack results on comments dataset. We draw the *Accuracy*, *Precision*, *Recall*, and *F1-Score* to compare the effectiveness of

selected features. Figure 4 *(red bars)* shows the performance of each classifiers by using selected features. According to the Figure 4 *(red bars)*, *Logistic Regression* performs better than other classifiers in all evaluation metrics. In the other hand, Figure 4 *(blue bars)* represent the reduction on *Accuracy* , *Precision* and *F1-Score* after hiding these 68 features from the comments feature set. As a result, we can infer the target user gender by using only comments dataset with the accuracy of 96%.

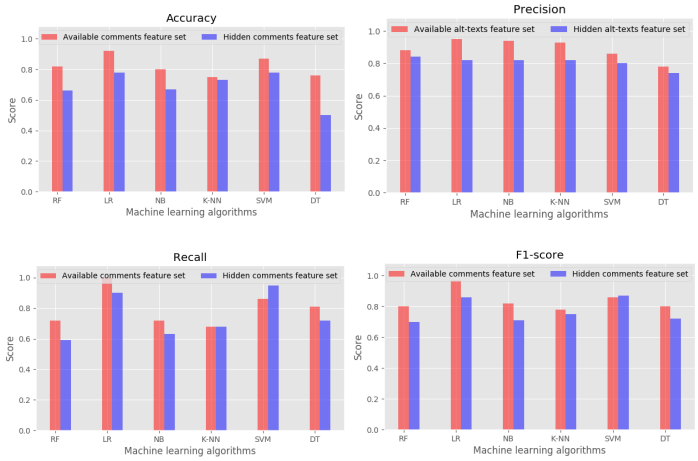


Fig. 4: Accuracy, Precision, Recall, and F-Score evaluation of comments dataset

5.4 Combined feature sets

In the last scenario, we train the machine learning classifiers by alt-texts and comments datasets. We union Algorithm 1 generated alt-texts and comments feature set. We refer to this set as a *combined feature set* which contain 134 features. We run Algorithm 1 on top of the combined feature set to check if there is a feature set that presents higher accuracy. With that, we create a new combined feature set by 97 features. *Univariate feature selection* perform the best on the combined feature set. Figure 5 *(red bars)* outlines the performance of each classifier by using the new combined feature set (97 features). Figure 5 *(red bars)* shows that the *Logistic Regression* perform better than other classifiers in *Accuracy*, *Recall*, and *F1-Score* by using these 97 features. In the other hand, Figure 5 *(blue bars)* shows the performance of classifiers by using those 134 features. It is observable that the classifiers metrics vary slightly. By the result, we understand that *Logistic Regression* performs best in our study. *Logistic Regression* is a discriminative model which is appropriate to conduct when the dependent variable is binary. So, it learns better between the dependent and independent variable in our dataset. According to the result, the combination of alt-texts and comments dataset gives the highest accuracy. That means, we can infer the gender of the target user by a high probability of 98% if both alt-texts and comments are available.

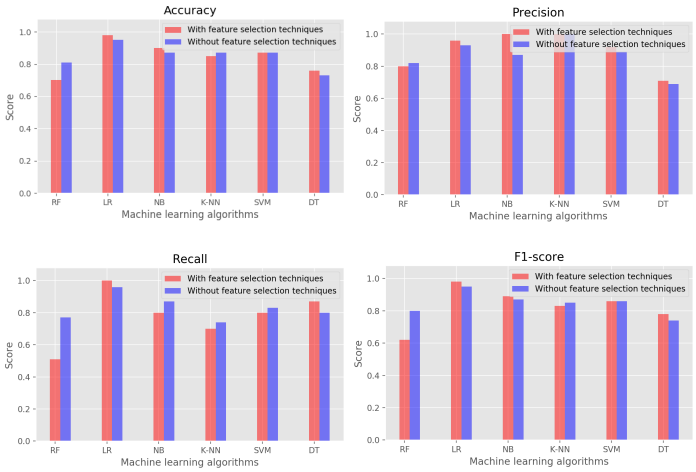


Fig. 5: Accuracy, Precision, Recall, and F-Score evaluation of combined dataset

6 Conclusion

In this work, we have presented a gender inference attack on Facebook by leveraging some easily collected datasets. Unlike other works, we have used *comments* that are written by target user friends, friends of friends or regular users and *alt-text* generated by Facebook to infer the target user gender. This type of attack can be categorized as an indirect attack. We show its possible to infer the target user gender even if the user hides his/her attributes and activities such as (profile attributes, friend list, liked pages and groups). In a nutshell, our results demonstrate that anyone can find hidden attributes of Facebook users they gender in their profile by just considering simple available information. As future work, we plan to propose counter-measures to picture owners, considering a trade-off between privacy risks and comments-based social benefits.

References

1. J. A. Bargh, K. Y. McKenna, and G. M. Fitzsimons. Can you see the real me? activation and expression of the true self on the internet. *Journal of social issues*, 58(1):33–48, 2002.
2. A. Chaabane, G. Acs, M. A. Kaafar, et al. You are what you like! information leakage through users interests. In *Proceedings of the 19th Annual Network & Distributed System Security Symposium (NDSS)*, 2012.
3. Y.-Y. Chen, T. Chen, W. H. Hsu, H.-Y. M. Liao, and S.-F. Chang. Predicting viewer affective comments based on image content in social media. In *proceedings of international conference on multimedia retrieval*, page 233. ACM, 2014.
4. K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

5. A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies*, 2015(1):92–112, 2015.
6. F. Fkih and M. N. Omri. Learning the size of the sliding window for the collocations extraction: A roc-based approach. In *Proceedings of the 2012 International Conference on Artificial Intelligence (ICAI12)*, pages 1071–1077, 2012.
7. N. Z. Gong and B. Liu. Attribute inference attacks in online social networks. *ACM Transactions on Privacy and Security (TOPS)*, 21(1):3, 2018.
8. N. Z. Gong, A. Talwalkar, L. Mackey, L. Huang, E. C. R. Shin, E. Stefanov, E. R. Shi, and D. Song. Joint link prediction and attribute inference using a social-attribute network. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(2):27, 2014.
9. J. He, W. W. Chu, and Z. V. Liu. Inferring privacy information from social networks. In *International Conference on Intelligence and Security Informatics*, pages 154–165. Springer, 2006.
10. R. Kang, S. Brown, and S. Kiesler. Why do people seek anonymity on the internet?: informing policy and design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2657–2666. ACM, 2013.
11. K. Lee. Combining multiple feature selection methods. In *Mid-Atlantic Student Workshop on Programming Languages and Systems (MASPLAS'02)*, pages 12–1. Citeseer, 2002.
12. J. A. B. Lopes Filho, R. Pasti, and L. N. de Castro. Gender classification of twitter data based on textual meta-attributes extraction. In *New Advances in Information Systems and Technologies*, pages 1025–1034. Springer, 2016.
13. M. Merler, L. Cao, and J. R. Smith. You are what you tweet pic! gender prediction based on semantic analysis of social media images. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2015.
14. H. Mihaljević and L. Santamaría. Telling the gender from a name, 2018.
15. Z. Miller, B. Dickinson, and W. Hu. Gender prediction on twitter using stream algorithms with n-gram character features. *International Journal of Intelligence Science*, 2(04):143, 2012.
16. A. Stavrianou, C. Brun, T. Silander, and C. Roux. Nlp-based feature extraction for automated tweet classification. *Interactions between Data Mining and Natural Language Processing*, 145, 2014.
17. A. Tobin and J. B. Merrill. Facebook is letting job advertisers target only men, 2018.
18. C.-F. Tsai and Y.-C. Hsiao. Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, 50(1):258–269, 2010.
19. U. Weinsberg, S. Bhagat, S. Ioannidis, and N. Taft. Blurme: Inferring and obfuscating user gender based on ratings. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 195–202. ACM, 2012.
20. S. Wu, J. Wieland, O. Farivar, and J. Schiller. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1180–1192. ACM, 2017.
21. W. Xu, X. Zhou, and L. Li. Inferring privacy information via social relations. In *2008 IEEE 24th International Conference on Data Engineering Workshop*, pages 525–530. IEEE, 2008.
22. J. Yurchisin, K. Watchravesringkan, and D. B. McCabe. An exploration of identity re-creation in the context of internet dating. *Social Behavior and Personality: an international journal*, 33(8):735–750, 2005.