# Your Installed Apps Reveal Your Gender and More!


## ABSTRACT

In this paper, we highlight a potential privacy threat in the current smartphone platforms, which allow any third party to collect a snapshot of installed applications without the user's consent. This can be exploited by third parties to infer various user attributes similar to what is done through tracking. We show that using only installed apps, user's gender, which is one of the common demographic attributes used in targeted advertising, can be instantly predicted with an accuracy around 70%, by training a classifier using established supervised learning techniques.


## 1. INTRODUCTION

Smartphone usage and the associated app market ecosystem are expanding rapidly. The current 25% − 30% smartphone penetration in the mobile device market is expected to reach 60% by the year 2019 [18]. The two most popular smartphone operating systems, Android and iOS, which have a combined market share of 91% [8], reached the one million mark in terms of number of apps in their official markets in 2013 [27, 10]. 32% of the app monetisations come through advertising [20] and in year 2014 mobile advertising revenue is expected to reach $31.45 billion, showing 75% increment [17] compared to previous year.

Increased app monetisation through advertising raises multiple privacy and security issues as developers try to collect user data through *over-permission* (i.e. asking for permissions which are not required for the functionality of the app) and share with third parties. This information is used to provide personlised advertisements, which have a higher likelihood of catching users' interest. For example, according to Yan et al. [28], behavioral targeted advertisements can significantly improve advertisement *click-through-rates*.

Information leakage through *over-permission* can be controlled by the users, as they can either control permissions an app has access to or uninstall the apps which ask for *over-permission*. However, in the two most popular smartphone platforms Android and iOS, the list of installed apps can be collected by any app developer or an embedded tracking library without the knowledge of user [12, 4]. This allows third parties to learn a range of information about the user instantly, which can be equivalent to inferences made after a period of data collection through tracking.

In this paper, we show how a third party can infer smartphone users' gender by mining a corpus of apps installed by users. First, we provide a basic characterisation of the differences in app installation patterns between male and female users, using a dataset collected from over 200 smartphone users. Then, we convert these differences into features and build a *linear SVM classifier* to predict a smartphone user's gender, given the list of installed applications and show that accuracy around 70% can be achieved. Finally, we compare our results with various other tracking-based data sources and show that the snapshot of installed apps performs comparable given the predictions are instant.

Our results highlight a privacy threat in the current smartphone ecosystems, as some users are not comfortable with disclosing their gender online. For example, according to Chaabne et al. [7], 20% of the *Facebook* users did not reveal their gender. Moreover, this type of instant inferences can be exploited by the service providers to quickly verify or assign a confidence value to user-entered information in smartphone apps. According to *Facebook* [19], 5.5%−11.2% of monthly active users of *Facebook* are fake users. Another study [16], showed that in the popular online game *World of Warcrafts*, 23% of the male users used female avatars, while 7% of the female used male avatars. There can be multiple reasons for users to give fake information about gender, with privacy concerns being one of them. Nevertheless, unrestricted access to other information such as installed app lists allows the service providers to identify users who entered fake information, which may be undesirable for some users.

The remainder of this paper is organised as follows. Section 2 discusses the related work. Section 3 describes our dataset. Section 4 presents the prediction methodology and performance of the predictions. Section 5 compares our results with prior work and Section 6 concludes the paper.

## 2. RELATED WORK

The possibility of demographic inference through various online footprints of the users has been discussed multiple times [25, 15, 11, 14, 24, 22, 21, 5, 30]. One such data source is the content and style of writing. Using a corpus of content obtained from 71,000 blogs at *blogger.com*, Schler et al. [24] showed that *writing style related features* such as parts-of-speech, function words, and hyper-links and *content-based features* such as simple content words and special classes of words taken from the handcrafted LIWC (Linguistic Inquiry and Word Count) [23] categories can be used to predict the writer's gender and age.

Similarly, Schwartz et al. [25] showed how the language in *Facebook* status update messages can be exploited to predict user demographics, age and gender as well as user's personality using *differential language analysis*. Otterbacher et al. [22] predicted the gender of the authors of IMDB movie reviews based on stylistic and content features.

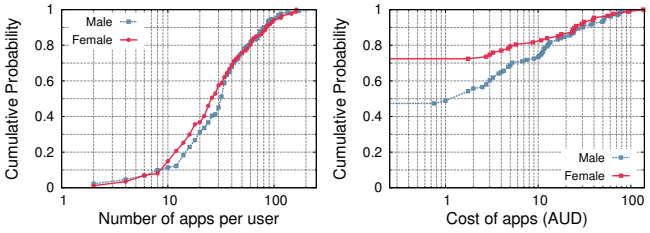
Another source of data used for demographic prediction is the web browsing patterns. Hu et al. [14] used data on page clicks from a major website to predict age and gender of the users, through a *Bayesian framework*. Goel et al. [11] used one-year client-side browsing history of 250,000 users to predict user's age, gender, race, education and income through *linear Support Vector Machines*.

Bi et al. [5] predicted age, gender, religion, and political views of the users based on search queries using models trained from *Facebook* likes. Ying et al. [30] used *behavioral features* in the likes of application usage, SMS usage, voice call usage, and *environment features* in the likes of number of Bluetooth and WiFi devices detected per day on mobile phones to predict demographics such as gender, age, and relationship status.

In prior work [**][1], we developed a framework to identify user traits such as languages spoken and relationship status through the presence of individual apps. However, this method does not work for gender as it cannot be directly associated to the presence of individual apps. To the best of our knowledge, this is the first attempt to predict smartphone users' gender, solely based on installed apps.

## 3. DATASET & FEATURE ENGINEERING

We collected a dataset of installed apps and the corresponding user demographic attributes through an Android app [**]. When installed, this app uploads the list of *user installed apps*[2] to a server and generates a random identifier. The app was distributed among volunteers and users recruited via Amazon Mechanical Turk [1]. These users installed the app and answered a brief questionnaire about demographic attributes such as gender and age together with the random identifiers so that the questionnaire output could be correlated with the collected app list.




(a) CDF: Number of apps (b) CDF: Cost of apps

Figure 1: Numeric features

Out of 218 users who provided an answer to gender, 131 ($\approx 60\%$) users were male and 87 users were female. In total, there were 4167 unique apps and on average each user had approximately 40 *user installed apps* in their smartphones. For each application, we queried official Android app market, Google Play Store and collected the metadata of the apps such as price, category and the app description. In the next subsections, we define and characterise various features that can be used to predict user's gender using list of installed apps.

### 3.1 Numeric Features

Figure 1a shows the *cumulative distribution function (CDF)* of the number of *user installed apps* for each gender. No significant differences can be found between the two genders as the two graphs closely follow each other.

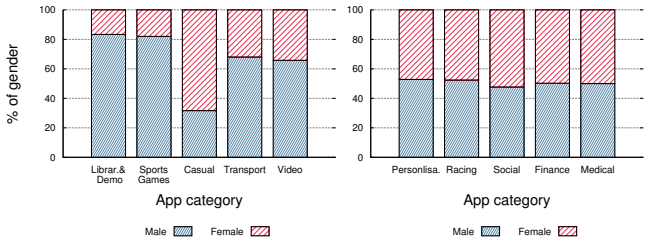
Figure 1b, shows the total cost of the *user installed apps*. We observe that male users tend to have more paid apps than female users. For example, approximately 70% of the female users had no paid apps installed while the corresponding value is only 50% for male users. However, when the money spent increases the difference in the gender distribution decreases. For example, the users who spent more than AUD$11.00, are equally likely to be male of female.

### 3.2 Category-based Features

Google Play Store categorises apps into 30 categories. For each app in our dataset, we queried Google Play Store and obtained the assigned app category. Then we tried to identify the categories that show a high difference in popularity between the two genders. We calculated the *discrete entropy* for each app category with respect to gender. Since we have more male users compared to female users, for the entropy analysis and for the succeeding subsections, we *under-sampled* the male users to match the number of female users.

Figure 2a, shows the 5 app categories with lowest en-

[1]** Citation is anonymised as the conference is double-blind.
[2]Android developer API provides a flag to differentiate *user installed apps* from *system installed apps* [2].




(a) High dispersion in gender (b) Low dispersion in gender

Figure 2: Gender dispersion in categories

| | $Adj(m^*) > Adj(f^*)$ | $Adj(f^*) > Adj(m^*)$ |
|---|---|---|
| Male | 80.92% | 14.50% |
| Female | 32.18% | 63.22% |

Table 1: Adjacency matrix ($m^*$ - Male and $f^*$ - Female)

tropy for gender (i.e. highest difference in popularity between genders) and Figure 2b shows the 5 app categories with highest entropy. As can be seen from Figure 2a, categories *Libraries & Demos* and *Sports Games* are more popular among male users while the category *Casual* is more popular among female users. Categories such as *Personlisation*, *Racing* and *Social* are equally popular between both the genders. To convert these observation to features related to individual users, we selected the percentage of apps in each category as the features representing individual users.

## 3.3 Item-based Features

Individual apps showing lowest entropies according to gender are shown in Figure 3. Since there were a large number of apps in our dataset, for this analysis we selected only the apps, that have been installed by at least 10% of the users.

Notice that certain apps show a strong tendency towards one gender. For example, apps such as *Google Translate* and *Reddit is fun* are more popular among male users while the apps such as *Pinterest* and *Ibotta-Cash* are more popular among female users. Some of these observations corroborate market reports published on app popularity among the genders. For example, according to [26] 83% of the users of the photo sharing app *Pinterest* are female, and 97% of the fans of *Pinterest's Facebook* page are female. According to [6] around 80% of the users of the game *Candy Crush Saga* are female. We selected the presence or absence of the top-10 apps showing the lowest entropy as features to represent individual users.

To represent the users in a more generalised manner, we calculated a *gender adjacency* for each user for the two genders following a similar approach to the Bayesian framework proposed by Hu et al. [14] for web pages. First, for each app, $a_i$ installed by more than 10% of the population, we calculated the probability of a user having that app being male or female, i.e. $p(c_i|a_i)$ where $c_i \in \{male, female\}$. Then we calculated *gender adjacency*, $Adj(c_i)$ for each user assuming the app installations are independent, i.e. $Adj(c_i) \propto \prod_{i=1}^{N} p(c_i|a_i)$, where $N$ is the number of overall popular apps (apps installed by more than 10% of the population) available in user's list of apps.

Table 1 shows the *adjacency matrix* for both female and male users. As can be seen, for approximately 81% of the male users, *male adjacency* is high compared to *female adjacency*. Only 19% male users had *female adjacency* higher than *male adjacency*. However, *female adjacency* is not a strong indicator for female users, as only 63% of the female users had *female adjacency* higher than *male adjacency*. Approximately 4.6% of users in each gender did not have any of the overall popular apps installed and for those users *gender adjacency* could not be calculated. We selected both *male adjacency* and *female adjacency* as features to represent users.

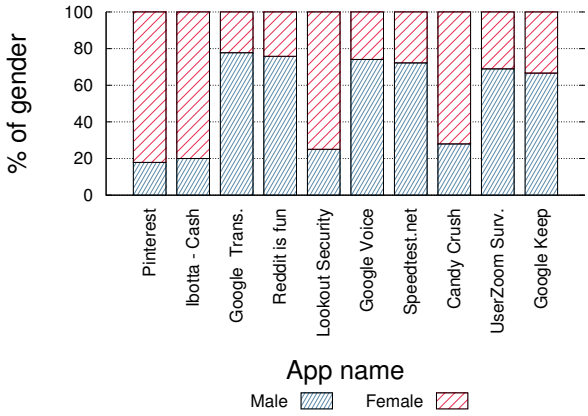



Figure 3: Gender distribution: Apps with lowest entropy values

## 3.4 Content-based Features

Google Play Store uses only 30 categories to categorise all the apps and therefore an app category can cover a broad range of topics. For example, app category *Lifestyle* can contain apps belonging to a range of topics such as real estate, religion, cookery or beauty care. We try to further narrow down these categories using content-based features.

For each app, we obtained the app description text provided by the app developer that usually explains the functionality of the app. We then represented each user as a concatenated text of the descriptions of all the apps installed on the phone. We selected only the apps with English app descriptions using the *Detect Language* API [3].

We then applied standard text mining techniques, specifically *stop word removal*, *change to lower case*, and *stemming* to preprocess the data. From the total collection of documents (one document for each user) we first identified the terms, which are present among at least 10% of the population. Then we selected top


Figure 4: Top-50 words with highest *information gain* with respect to gender

$n$ terms, $n \in \{100, 1000\}$, when the terms are ranked according to the value of *information gain* [29].

We calculated the *tf-idf* weights for the selected terms and used those as features to represent each user. Figure 4, visualises the top-50 words having the highest *information gain* with size being proportional to the value of *information gain*. As the figure illustrates, there are terms which are only popular among users in one gender. For example, terms *network* and *rss* are popular among male users, whereas terms *decor* and *pregnanc* are more popular among female users.

## 4. METHODOLOGY & RESULTS

From the total dataset, we randomly selected 50% of user data for training and 50% for testing. To measure an average classifier performance, we built 5 training and testing set pairs by sampling the original dataset. Then, we represented each user as a feature vector using the features mentioned in Sections 3.1 to 3.4. To address class imbalance, we randomly under sampled the majority male user class to match the size of the female user class in the training set.

We trained naive Bayes and linear kernel SVM (*Support Vector Machine*) classifiers [9] using the machine learning tool Weka [13] using individual categories of features as well as feature categories in combination. Table 2 summarises the performance of the classifiers observed over the 5 samples.

The results show that numeric features alone do not perform better than the random classification, as the AUC (*Area Under the Curve*) values are approximately around 0.5. Category-based features show higher performance under the SVM model achieving 72% accuracy and 69% AUC values.

Item-based features perform better compared to random prediction; however, they do not outperform category-based features. From the content-based features, selection of more terms (1K) performed better compared to the selection of lower number of terms. Combining

Table 2: Performance of the classifiers
NB* - Naive Bayes SVM* - Support Vector Machines

| Feature type | Classi. | Accuracy | AUC |
|---|---|---|---|
| Numeric | NB | 0.493 | 0.499 |
| | SVM | 0.602 | 0.500 |
| Category | NB | 0.618 | 0.689 |
| | SVM | **0.715** | **0.693** |
| Item | NB | 0.635 | 0.695 |
| | SVM | 0.679 | 0.640 |
| Content (100) | NB | 0.663 | 0.722 |
| | SVM | 0.641 | 0.646 |
| Content (1K) | NB | **0.687** | **0.740** |
| | SVM | 0.680 | 0.686 |
| Category+Item | NB | 0.650 | 0.698 |
| | SVM | 0.683 | 0.667 |
| Category+Content (1K) | NB | **0.698** | **0.740** |
| | SVM | 0.681 | 0.689 |
| Category+Item+Content (1K) | NB | **0.698** | **0.737** |
| | SVM | 0.685 | 0.690 |

Table 3: Classifier Performance
$Category + Content(1K)$

| Sample | Gender | Precision | Recall | AUC | Accuracy |
|---|---|---|---|---|---|
| 1 | Male | 0.817 | 0.754 | 0.784 | 0.750 |
| | Female | 0.667 | 0.744 | 0.76 | |
| 2 | Male | 0.727 | 0.615 | 0.681 | 0.630 |
| | Female | 0.528 | 0.651 | 0.666 | |
| 3 | Male | 0.849 | 0.692 | 0.802 | 0.741 |
| | Female | 0.636 | 0.814 | 0.756 | |
| 4 | Male | 0.733 | 0.704 | 0.776 | 0.704 |
| | Female | 0.596 | 0.791 | 0.728 | |
| 5 | Male | 0.704 | 0.769 | 0.674 | 0.667 |
| | Female | 0.595 | 0.512 | 0.709 | |

features did not provide a significant performance improvement. For example, combining other feature categories with content-based features (1K) did not improve the performance significantly; indicating content-based features are the most influential features deciding the classifier performance.

To explore further, we selected the naive Bayes classifier built using *Content (1K)+Category* features as the best performing classifier as it has the highest classifier performance and accuracy and analysed its performance in detail. Table 3, shows the accuracy, precision, recall, and the AUC of the selected classifier over the 5 samples.

According to Table 3, 3 samples showed over 70% accuracy. Across all samples a precision over 70% could be achieved for male user prediction with a recall varying between $62\% - 77\%$. Precision for female users varied between $53\% - 67\%$ with a recall varying between $51\% - 81\%$. The performance of the classifiers can be considered as moderate, as they perform better compared to the random classification, and do not show very high accuracy.

## 5. DISCUSSION

We showed using a single snapshot of installed apps; gender of smartphone users can be predicted with an accuracy around 70%. In this Section, we compare our results with other online data sources that can be used for predicting the gender of the users.

In Table 4, we compare our results with the work of Hu et al. [14] in terms of precision and recall. Performance of our classifier was lower than the classifier built based on users' web page clicks. However, prediction of male users showed a close precision value only at a 10%

Table 4: Performance comparison
Apps ($Category + Content - 1K$) vs. web page clicks

| Related Work | Gender | Precision | Recall |
|---|---|---|---|
| Wep page click logs [14] | Male | 0.791 | 0.810 |
| | Female | 0.805 | 0.782 |
| Apps - *Category+Content (1K)* | Male | 0.767 | 0.701 |
| | Female | 0.604 | 0.702 |

Table 5: Performance comparison
Apps ($Category + Content - 1K$) vs. other data
(NA* - Result of the performance metric not available)

| Related Work | Accuracy | AUC |
|---|---|---|
| *Facebook* likes [15] | NA* | 0.93 |
| *Facebook* status updates [25] | 91.9% | NA* |
| Search queries [5] | NA* | 0.803 |
| Client side browser history [11] | $\approx$ 75% | $\approx$ 0.85 |
| Mobile phone usage & environment [30] | $\approx$ 82-85% | NA* |
| Apps - *Category+Content (1K)* | 70% | 0.74 |

drop in recall. Hu et al. used a one-week long dataset, whereas we used only a single snapshot of installed apps.

Table 5, compares our work with other work, which used *Facebook* likes [15], language style in *Facebook* status updates [25], search queries [5], client side web browsing logs [11] and mobile phone usage monitoring [30]. Our method performed close to the use of client side browser history in terms of accuracy. Further the performance is close to the performance of search query and mobile usage monitoring. A difference of approximately $12\% - 15\%$ in accuracy with the mobile usage monitoring method is interesting as monitoring usage activities on a mobile phone is a resource intensive task and considered as violating user privacy.

Comparisons show that in the smartphone context, the use of a snapshot of installed apps to predict gender provides performance comparable to other data sources and methods that require much richer datasets collected by tracking user activities over a period of time or collecting more intrusive personal data such as *Facebook* likes or status updates. We believe access to a much larger dataset of the users' installed apps and the corresponding ground truth of gender will enable higher classifier performance as more advanced features can be generated. Moreover, it might be possible to infer other user attributes such as age, ethnicity, and income level following a similar method.

# 6. CONCLUSION

This paper provided insights on various features related to installed apps on a smartphone, which can be used to predict the gender of the user. Using a dataset of over 200 smartphone users, we showed that by only observing a single snapshot of installed apps, it is possible to predict gender of the smartphone user with an average accuracy of approximately 70%. We compared our results with other data sources used for gender prediction, which involve data collection for longer periods and/or require more personal data from users such as *Facebook* likes or their social media status updates. We showed that through a single snapshot of apps we could achieve comparable results to longer and more intrusive inference techniques.

# 7. REFERENCES


[1] Amazon mechanical turk. https://www.mturk.com/.
[2] Applicationinfo: Android Developers. http://developer.android.com, 2013.
[3] Language Detection API. http://detectlanguage.com, 2013.
[4] D. Amitay. iOS App Detection. http://www.ihasapp.com, 2012.
[5] B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel. Inferring the demographics of search users: social data meets search queries. In *Proc. of the 22nd WWW*. ACM, 2013.
[6] D. Bolton. Why I stopped playing Candy Crush Saga. http://news.dice.com, 2013.
[7] A. Chaabane, G. Acs, M. A. Kaafar, et al. You are what you like! information leakage through users' interests. In *Proc. of the 19th NDSS*. The Internet Society, 2012.
[8] A. Cocotas. Android grabs a record share of the global smartphone market. http://au.businessinsider.com, 2013.
[9] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3), 1995.
[10] S. Fiegerman. Apple's App Store tops 1 million apps. http://mashable.com, 2013.
[11] S. Goel, J. M. Hofman, and M. I. Sirer. Who does what on the web: A large-scale study of browsing behavior. In *Proc. of the 6th ICWSM*, 2012.
[12] M. Grace, W. Zhou, X. Jiang, and A. . Sadeghi. Unsafe exposure analysis of mobile in-app advertisements. In *Proc. of the 5th WiSec*, pages 101–112. ACM, 2012.
[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.
[14] J. Hu, H. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *Proc. of the 16th WWW*, 2007.
[15] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proc. of the National Academy of Sciences*, 110(15):5802–5805, 2013.
[16] R. M. Martey, J. Stromer-Galley, J. Banks, J. Wu, and M. Consalvo. The strategic female: gender-switching and player behavior in online games. *Information, Communication & Society*, pages 1–15, 2014.
[17] eMarketer Inc. Driven by Facebook and Google, mobile ad market soars 105% in 2013. http://www.emarketer.com, 2014.
[18] Ericsson Inc. Ericsson Mobility Report. http://www.ericsson.com, 2013.
[19] Facebook Inc. 2013 Annual Report. http://investor.fb.com/annuals.cfm, 2013.
[20] M. Meeker. Internet Trends 2014. http://www.kpcb.com/internet-trends, 2014.
[21] A. Mukherjee and B. Liu. Improving gender classification of blog authors. In *Proc. of the 7th EMNLP*, pages 207–217. Association for Computational Linguistics, 2010.
[22] J. Otterbacher. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proc. of the 19th CKIM*. ACM, 2010.
[23] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71, 2001.
[24] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205, 2006.
[25] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9), 2013.
[26] K. Wagstaff. Men are from Google+, women are from Pinterest. http://techland.time.com, 2012.
[27] C. Warren. Google Play hits 1 million apps. http://mashable.com, 2013.
[28] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen. How much can behavioral targeting help online advertising? In *Proc. of the 18th WWW*, pages 261–270. ACM, 2009.
[29] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420, 1997.
[30] J. J. Ying, Y. Chang, C. Huang, and V. S. Tseng. Demographic prediction based on users mobile behaviors. In *Proc. of the MDC*. Nokia, 2012.