Arikui - A Dubious User Detection System for Online Dating in Japan

James Neve, *Eureka Inc., Tokyo, Japan*
Ivan Palomares Carrascosa, *Univeristy of Bristol, Bristol, United Kingdom*

***Abstract*—Online dating constitutes one out of myriad popular services that can be accessed via the Internet nowadays. This paper introduces a novel detection system for identifying dubious users, i.e. users who utilize a Japanese online dating service for purposes besides dating. Examples of such purposes include sales and multi-level marketing, amongst others. More specifically, the proposed detection is characterized by simultaneously analyzing: (i) user profile data; (ii) user actions over their first few hours; and (iii) data retrieved from Facebook in order to find the likelihood that the user is a spammer. The resulting system successfully detects a number of spammers every day, thereby becoming a valuable tool for the customer service team in Eureka Inc, where it has been deployed.**

***Index Terms*—spam detection, machine learning, information system, big data**

I. Introduction

The online dating market in Japan differs drastically from those found across other parts of the globe. Online dating is regarded with great suspicion by the older generations and part of the younger generations alike[4]. Traditionally, relationships between young people in Japan are initiated at *Goukon* -parties where friends introduce mutual friends to each other. For people closer to and over 30 years old, remaining unmarried may become less socially acceptable, depending on the specific cultural and social background. In such situations, *Konkatsu* services, which aim at kickstarting potential marriage partnerships personally vetted by the service, are often utilized[6].

With the advent of smartphones, the vast majority of Japanese apps related to dating present the - often not desirable - characteristic of having large numbers of fake users intended to make money from real users. Only in the last five years have the younger generation begun to see online dating as a viable method to find a serious relationship, and several services have emerged to supply this demand[12]. It is therefore vital that, when a Japanese person with a natural bias against online dating services decides to try them, their a priori misconceptions are not are not immediately confirmed. Changing such cultural misconceptions implies ensuring that serious dating services are properly monitored predicated on as many informational resources as possible.

Most existing dating services are used for non-legitimate purposes by a certain percentage of users. The three most notorious categories of "suspicious users" are listed as follows.

1) *People selling as part of a company*. They will usually try to arrange to meet other people for a date, and then make the sale after meeting. They are often serial users of community sites, who know how to write profiles and messages similarly as a real user would normally do. Often, these dubious users will have a large number of apparently active social media accounts that they can use for registration. They are almost always masquerading as visually attractive persons.
2) *Multilevel marketing or other fraudulent schemes*[11]. They will act in a similar way to the above mentioned users, but will generally use their personal accounts to register, and won't act with the same level of professionalism, sometimes trying to sell to other users within their first two or three messages.
3) *Members looking for a quick "hook-up" rather than a real relationship*. Almost always male users, they will usually send some kind of rude, tactless or inappropriate message within the first one or two interactions.

Simple word and pattern filters are in most cases sufficient to detect the third type of dubious user. However, the first two types of dubious user are much more difficult to catch. In particular, if a honest user attends a 'date' that ends up being a sales pitch, then (s)he would become much more likely to leave the service. Dubious users are aware that they will be removed if they are found, and so whether they are amateur or professional, they attempt to mimic normal users as closely as possible. This makes them remarkably difficult to detect and deal with by using simple filtering techniques as mentioned above.

Recent advances in machine learning have undeniably made it an effective tool for pattern recognition in continuous data. In particular, "Deep Learning" has been used effectively for classifying images, sounds and videos [16]. Intuitively, similar techniques can be used to handle dubious users in categories (1) and (2) by incorporating temporal data describing their activity and behavior. In this way, despite their attempts to masquerade as normal users from other users' points of view, deviations from normal behavior are nonetheless shown when examined under a temporal dimension. For example, they are more likely to (i) be very active immediately after joining, (ii) more likely to send similar messages to every user they interact with and (iii) more likely to be active during normal working hours.

Eureka[1], a popular online dating services provider, has access to a large data set of approximately six million normal (legitimate) and dubious users from the dating app *Pairs*, build up over five years. *Pairs*[2] users regularly report interactions

[1]https://eure.jp/en/about/
[2]https://pairs.lv/

with all categories of dubious users to Eureka's customer service team, who can subsequently flag such users. Accordingly, in this paper we introduce a dubious user detection system characterized by enabling a rapid, accurate detection response based on behavioral and profile data. The main innovation in this research, which has been validated and deployed in the *Pairs* app, relies in the efficient and successful detection of dubious users in the context of online dating, *before* they are able to exchange contact details with normal users. Both user profile data, user behavior data (particularly during the earliest stage of using the app) and external social media data e.g. from Faceboook, are analyzed.

In summary, the main characteristic of the dubious user detection system are listed as follows:

1) It detects dubious users as described above in categories (1) and (2).
2) It enables an effective early-stage detection within 24 hours of registration, before such dubious users have had time to send marketing messages, exchange contact details or set up meetings with a large number of normal users.
3) It reduces the number of false positives (legitimate users categorized as dubious) in the detection results, so as to minimize the need for unnecessary checking by the customer service department.

The remainder of this paper is set out as follows. Section II provides a summary of the literature on similar works. Section III describes the architecture of the dubious user detection system. Section IV describes how the system operates in practice. Section V provides and summarizes some experimental results. Finally, the achievements and areas for further work are described in Section VI.

II. Related Work

There has been a considerable deal of research over the last few years devoted to Neural Networks in general and Deep Learning in particular. Time series analysis has been shown to greatly benefit from convolutional neural network models, with a large number of successful applications being developed in this area [10].

A thorough coverage of machine-learning driven financial fraud detection has been undertaken in the literature. In particular, the behaviour-driven approach has gained significant popularity, with a number of applications of dividing user actions into categories and blocks of time, and training convolutional networks on the resulting matrix [7]. *GOTCHA!* describes a system for finding fraud detection in social networks by modelling the network as a bipartite graph[13] and other systems that work through clique detection[15]. However, online dating has a somewhat unusual data model whereby users have contact with potential partners only, and dubious users do not form connections with each other. These approaches are therefore of limited use in this domain.

A related field to dubious user detection in online dating, is message and email spam detection. Beginning with simple word filters, and moving quickly on to simple statistical approaches such as Bayesian filters trained on individual word probabilities [14] and progressing to more complicated approaches such as Random Forest classifiers [5]. Although *Pairs* does employ a machine-learning based spam filter, it is of limited use in finding suspicious users. This is because written Japanese (and especially casual Japanese written in a variety of dialects) is extremely difficult to tokenize compared to English, and because, as discussed earlier, dubious users will often mimic ordinary users in their messages, and only attempt to sell products (thus acting illegitimately) once the meeting has taken place.

III. System Overview

This section provides on overview of the dubious user detection system integrated in the *Pairs* app and introduced in this paper. In the sequel, the data analysis steps underlying the detection process are described in detail. The process comprises three major steps: (i) a data analysis and transformation step, (ii) a profile filter and (iii) a behavior filter. Details of the implementations within each step are given in the next three subsections.

A. Data Analysis and Transformation

The dataset contains a large amount of user profile data. This includes continuous profile data such as age and height[3], discrete profile data such as whether they were immediately looking to marry, as well as free text profile data. Our initial expectation - and one of our research hypotheses - considers that dubious users would create profiles that look as attractive as possible to potentially target users, in order to increase the chances of getting a response.

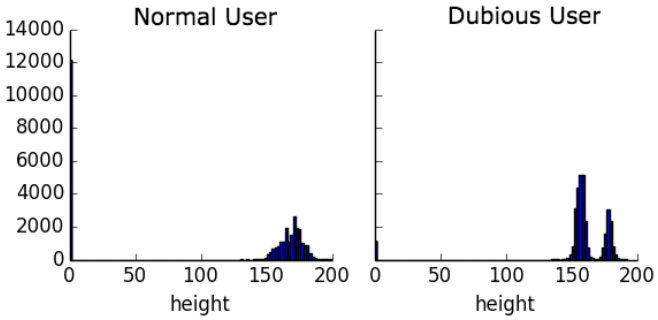

Fig. 1. Height analysis for normal and suspicious users

In line of such a notion of attractiveness in user profile data, statistical analyses of continuous data showed that dubious users are likely to be very close to the national average if they were female. Likewise, they are significantly taller than average if they were male, whereas ordinary users resemble a normal distribution much more closely. Age follows a similar trend, with ordinary users normally distributed, and dubious users reliably much younger than usual. A height or age close to the average obviously is not deemed as a sufficient condition to mark someone as a dubious user, as many normal

[3]Within the scope of the proposed detection system, age and height data take a wide range of possible values across users, and such data attributes are thus bucketed and treated as "continuous".

users share similar stats, but if those stats are further from the average, the likelihood of being dubious would be lower. In order to make the data discrete and reduce the impact of random variation, we bucketed discrete data.

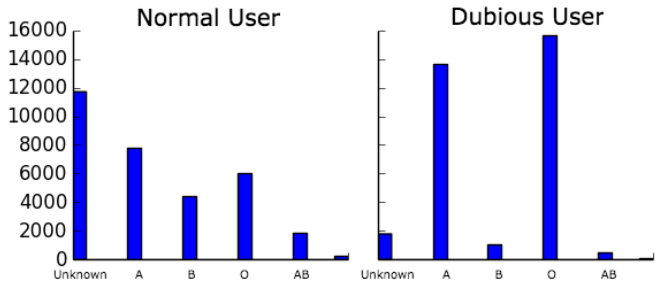

Fig. 2. Blood type analysis for normal and suspicious users

Discrete data inputted by users on their profiles was less significant in the detection process. Although there is a particular incentive for dubious users to make their profiles attractive, many ordinary users are also likely to bend the truth about their opinions or income in order to make their profiles more attractive to the opposite sex, so there are very few features where significant useful deviation appears. Two features where exhibitting a large such deviation, however, are smoking and blood types. Dubious users are almost exclusively non-smokers, which complies with the principle that dubious users try to seem attractive - smokers are much less attractive to non-smokers. The difference in blood type is a Japanese cultural phenomenon, where it is is widely believed to be a reliable predictor of personality. Dubious users were reliably *A* or *O*, whereas normal users had distributions close to the real population distributions.

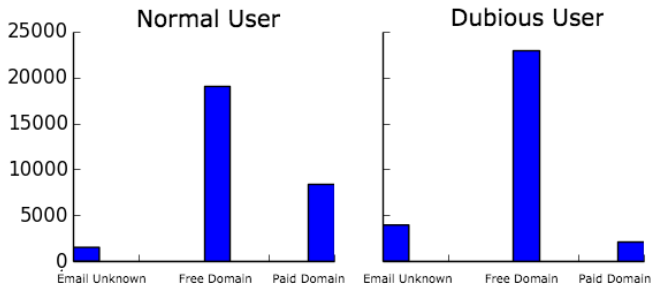

Fig. 3. Email domain analysis for normal and suspicious users

Free text data was the most difficult to extract meaningful differentiating information from. Unsurprisingly, normal users in an online dating service try to make their text profiles attractive similarly as dubious users would do. Hence, the challenges of natural language processing and tokenizing Japanese language both make extracting numerical data for machine learning difficult. Nonetheless, some meaningful information can be extracted from textual data. A simple analysis of text length showed that dubious user profiles tended to be significantly longer than others. Dubious user messages are usually very similar to each other, whereas normal users are more likely to tailor their messages to the user they were contacting, such that a simple Levenshtein[9] comparison between the first message and subsequent messages gave a good indication. Finally, dubious users are more likely to have less popular email domains.

B. Profile Filter

After classifying the existing user profile data, we now describe the process followed by the profile filter. As part of the 'profile filter', both profile data and data gathered from social media was used as input. Classified data comprise a number of discrete data points divided into normal users and dubious users, as identified by the customer service team. At the time the system was being developed, data for approximately 30,000 dubious users was available. 30,000 normal users were therefore extracted from the database. The total training dataset comprised 60,000 data points.

The following characteristics underlying the data influenced the choice of algorithm for the profile filter:

1) The algorithm chosen needs to converge relatively quickly. 30,000 dubious users were identified over the course of five years, so there were few opportunities for significantly increasing the size of the dataset in the immediate future. Algorithms or parameters that would require millions of datapoints to converge are not feasible options for this reason.
2) All of the fields chosen highly independent from each other. Therefore, algorithms that rely on examining relationships between contiguous data such as convolutional networks are not suitable in this context.
3) The quality of the data is not guaranteed. Dubious user data by the time the system was devised, was explicitly categorised as such by members of the customer service department, so the users in that dataset were guaranteed to be dubious. However, some dubious users end up not being reported by by other users, and therefore unexamined by customer service team members. It is possible, therefore, that a percentage of users in the 'normal' dataset are indeed real dubious users.
4) The data is not complete. Many users do not complete their entire profiles, but leave certain sections out. These sections were assigned a separate 'incomplete' value.

The data was split into 90% training data set and 10% test data set for cross-validation. The data was classified using a random forest algorithm. A random forest algorithm samples K features from the input dataset [3]. Decision tree classifiers are then trained on the basis of those randomly selected input vectors using bootstrap aggregation, which fits trees to random sets of the training data. The output for unseen data samples is then determined based on the mode of the results of the individual tree classifiers. Trees containing features that are strong predictors of the output are very likely to be strongly correlated, and will therefore increase the accuracy of the mode as a predictor.

For the profile, we used a random forest algorithm with 100 estimators. Based on cross-validation, it achieved 94.1% accuracy.

C. Behaviour Filter

Conceptually, the behavior filter is slightly different to the profile filter. Temporal information is available about user actions and the time at which they took place. The relationships between those timestamps is significant in the detection process. Dubious users have an incentive to connect with as many users as possible before they are detected and blocked from the site. Visual examinations of the data indicated that dubious users are much more proactive from the earliest stage of use of the application. They are more likely to 'like' other users, and more likely to initiate conversations.

In order to capture the temporal relationship between the data, a convolutional neural network is used to filter users according to their behavior over time. Convolutional neural networks have been particularly successful in image classification tasks[8], partly due to advances in hardware that allow very deep networks to be trained quickly. Convolutional networks generally consist of alternating layers of convolution layers (neurons connected to small regions of the input), usually with multiple filters per layer to create an output volume, and pooling layers, which downsample the results of the convolution filters. The final layers are generally fully connected layers that classify the convolution results.[2]

Convolutional neural networks have proved effective on handling temporal information in a number of cases[1]. The user behavior data is modeled as 3-dimensional array, containing hourly information about the number of profile updates, 'likes' (expressions of interest of preference towards other users' profile) and message exchange. The most effective type of network to deal with such data consists of two layers of convolution filters, each followed by a max-pooling layer, and a fully-connected layer of 1024 nodes at the end. The nodes were rectified linear units, and the network was trained to minimize cross entropy.

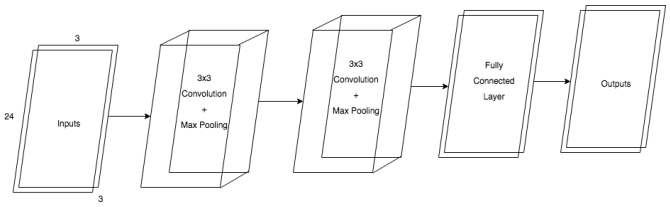

Fig. 4. Network architecture used to detect dubious users based on behaviour

By the time the behavioral filter was integrated with the overall detection system, a 96.2% accuracy was achieved based on cross validation, which is slightly higher than the profile filter. Based on visual inspection of the data, the behaviour of men and women is significantly different, and interestingly the behaviour of dubious women and normal men is relatively similar. Ideally, two different detection models would therefore have been created - one for men and one for women. However, as neural networks are relatively slow to converge compared to other machine learning algorithms, and considering that there exist a relatively small number of data points (30000 - 27000 for training and 3000 for cross-validation), creating two separate models would not improve the overall accuracy.

IV. System Operation

This section describes the infrastructure of the detection system in practice. Both the profile filter and the behaviour filter were build as Python models, and deployed to an EC2 instance. A microservice has been developed to send data to the model, and retrieve predictions. The customer service team at Eureka has access to a control panel, from which they can monitor suspicious user alerts and mark them as dubious or normal.

The profile filter was particularly challenging to implement effectively. As users can complete their profiles in any order, and any time after registration, users had to be re-evaluated each time their profile changed. Eventually, an effective system was settled on whereby customer service representatives were alerted if a user's suspiciousness was above a certain threshold, or if a user who had previously been marked as suspicious was marked with a higher score than they had been before based on a subsequent profile update.

The behaviour filter was run as part of a batch task, evaluating users who had been registered for more than 24 hours. Ideally, the behaviour filter would have been run based on more than 24 hours of registration, and would probably have predicted results more accurately. However, in practice most dubious users are reported within a few days of registration, and will try to exchange phone numbers, email addresses or other contact details as quickly as possible so as to remain in contact with users after being banned from the site. It is therefore vital to detect them before they have enough time to connect with individual users.

V. Performance Results in an Online Dating App

This section describes the results of the system operating on Pairs, and explains possible reasons for these results

Both the profile and behaviour filters achieved a high level of accuracy based on cross validation performed with an equal number of normal and dubious users. However, in practice, the systems reported a large number of false positives (i.e. normal users reported as suspicious) alongside truly suspicious users. This is a result of the relative population sizes of normal and dubious users.

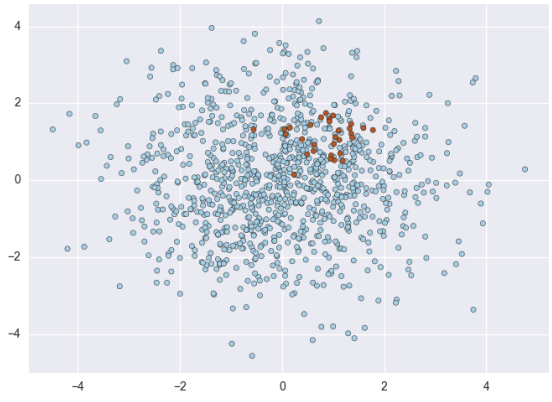

Fig. 5. Relative population sizes of normal and dubious users

Threshold	Marked Users	Dubious Users	Accuracy
0.995	287	43	15%
0.997	175	33	18%
0.998	106	24	23%
0.999	41	10	24%
0.9995	21	10	48%
0.9997	8	5	62%

TABLE I
RELATIVE POPULATION SIZES OF NORMAL AND DUBIOUS USERS

Although the training populations were equal, in reality normal users make up the vast majority of the population of users. There is no feasible approach to know exactly what proportion of dubious users are in the population, as some undoubtedly go undetected. However, identified and blocked suspicious users make up 0.8% of total users.

As a result of this, unless very high threshold values are set on both filters, false positives are more common than accurately identified dubious users. Pairs, as the most popular dating service in Japan, has on average between 100 and 300 registrations every hour depending on the time of day, which significantly increases at certain times of year (for example, New Year). A system that forces the customer service team to necessarily analyze a dozen users every hour, many of which are false positives, is not suitable in practice. Results for the profile filter based on week-long measurements are as follows.

Based on week-long trials at various thresholds, the accuracy of the system at various levels could be measured. In practice, a threshold accuracy value of 0.997 is used for validating the system. This intuitively seems extremely high, and a large number of dubious users are undoubtedly missed during the normal operation. However, the amount of human intervention required to check hundreds of users at a lower accuracy rate was reported as being not significant enough to justify a lower threshold.

In the 6 months since the system's inception, 854443 users have registered on the site. Of those, the system has correctly identified 12575 as spammers (0.015%). It is impossible to know how many total spammers there are, as the ones who have gone unidentified are not recorded. However, spammers will often contact a large number of users before being detected, so it is certain that in absolute terms, tens of thousands of spam messages or meetings were prevented.

VI. CONCLUSIONS

This contribution presented a system to detect dubious users on an online dating service, generalisable to other services.

A random forest-based classifier and a neural network-based classifier were developed, which in conjunction enable a highly effective detection of dubious users with a high degree of accuracy. In this sense, the system successfully achieves a desirable level of detection accuracy. In a balanced population, both the profile filter and the behaviour filter detect dubious users with approximately 95% accuracy: the filters are highly successful at differentiating between the two types of user.

However, considering the relative sizes of the populations of normal and dubious users, the system with a threshold set high enough that a large number of hours are not wasted checking false positives is also likely to miss a large number of real dubious users. This exemplifies the difficulty of using machine learning in a setting with very unbalanced populations. Many machine learning systems that approach a very high degree of accuracy in a setting with a balanced population can still output a large number of false positives where the task is to identify members of a small minority within the population.

Despite the above drawbacks, the system is a valuable tool for the Eureka customer service department, and and has effectively facilitated the identification and subsequent blockage of a large number of dubious users before they were able to contact a significant number of normal users.

There are various methods the system could be improved, constituting our most immediate directions of future work. In particular, the accuracy would probably increase significantly with larger training sets. 30,000 examples is relatively small training set in the context of machine learning. A larger training set also opens up the possibility of other models. For example, the dating site allows users to submit photos of themselves. Dubious users often pose as young, attractive women, and their photos are therefore likely to have certain characteristics that average users do not. However, a training set of 30,000 items is too small to build a neural network that gives useful results from image data. Theoretically, looking at behaviour over a longer period of time than 24 hours would also give more information to evaluate, and would likely improve the accuracy of the predictions, despite its non-guaranteed practicality from a customer service point of view.

REFERENCES

[1] S. Shetty A. Karpathy, G. Toderici. Large-scale video classification with convolutional neural networks. 2014.
[2] G. Hinton A. Krizhevsky, I. Sutskever. Imagenet classification with deep convolutional neural networks. 2012.
[3] M. Wiener A. Liaw. Classification and regression by randomforest. 2002.
[4] N. Li B. Hogan, W. Dutton. A global shift in the social relationships of networked individuals: Meeting and dating online comes of age. 2011.
[5] H. Wechsler D. DeBarr. Spam detection using clustering, random forests, and active learning. 2009.
[6] L. Dales E. Dalton. Online konkatsu and the gendered ideals of marriage in contemporary japan. 2016.
[7] Y. Tu K. Fu, D. Cheng. ?credit Card Fraud Detection Using Convolutional Neural Networks. 2016.
[8] A. Zisserman K. Simonyan. Very deep convolutional networks for large-scale image recognition. 2015.
[9] I. Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. 1966.
[10] A. Loutfi M. Lngkvist, L. Karlsson. A review of unsupervised feature learning and deep learning for time-series modeling. 2014.
[11] A. Rege. What's love got to do with it? Exploring online dating scams and identity fraud. 2009.
[12] T. Tsuruki TJM. Holden. Deiai-kei: Japan's new culture of encounter. in Gottlieb N, McLelland M, eds. Japanese cybercultures. 2003.
[13] L. Akoglu V. Van Vlasselaer, T. Eliassi-Rad. ?gotcha! network-based fraud detection for social security fraud. 2014.
[14] T. Eliassi-Rad V. Van Vlasselaer, L. Akoglu. An evaluation of naive bayesian anti-spam filtering. 2000.
[15] T. Eliassi-Rad V. Van Vlasselaer, L. Akoglu. Guilt-by-constellation: Fraud detection by suspicious clique memberships. 2014.
[16] G. Hinton Y. LeCun, Y. Bengio. Deep Learning. 2015.