International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 3, Juni 2020, pp. 2763~2772
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i3.pp2763-2772

Classification of instagram fake users using supervised machine learning algorithms

Kristo Radion Purba, David Asirvatham, Raja Kumar Murugesan
School of Computing and IT, Taylor's University, Malaysia

Article Info

Article history:

Received Oct 23, 2019
Revised Nov 26, 2019
Accepted Dec 9, 2019

Keywords:

Classification algorithm
Fake user
Machine learning
Paid follower
Social media

ABSTRACT

On Instagram, the number of followers is a common success indicator. Hence, followers selling services become a huge part of the market. Influencers become bombarded with fake followers and this causes a business owner to pay more than they should for a brand endorsement. Identifying fake followers becomes important to determine the authenticity of an influencer. This research aims to identify fake users' behavior and proposes supervised machine learning models to classify authentic and fake users. The dataset contains fake users bought from various sources, and authentic users. There are 17 features used, based on these sources: 6 metadata, 3 media info, 2 engagement, 2 media tags, 4 media similarity Five machine learning algorithms will be tested. Three different approaches of classification are proposed, i.e. classification to 2-classes and 4-classes, and classification with metadata. Random forest algorithm produces the highest accuracy for the 2-classes (authentic, fake) and 4-classes (authentic, active fake user, inactive fake user, spammer) classification, with accuracy up to 91.76%. The result also shows that the five metadata variables, i.e. number of posts, followers, biography length, following, and link availability are the biggest predictors for the users class. Additionally, descriptive statistics results reveal noticeable differences between fake and authentic users.

Corresponding Author:

Kristo Radion Purba,
School of Computing and IT,
Taylor's University,
1, Jalan Taylors, 47500 Subang Jaya, Selangor, Malaysia.
Email: kristoradionpurba@sd.taylors.edu.my

1. INTRODUCTION

Instagram is the third most used social media in terms of the number of active users [1]. It is also the most popular social media for teens [2], as well as influencer marketing [3]. Commonly, social media has like, comment, and follow functions. The success of an influencer is commonly measured by these numbers [4]. In this regard, fake followers become important for them. The number of followers and likes are commonly perceived as a social status [5]. Commonly, fake users refer to bots. However, a report found that it can also include users who sell their passwords [6]. Fake users are a huge concern for Instagram, where there are an estimated 10 million such users [7]. Fake users also create a problem for business owners who pay the influencer for endorsement. This endorsement cost is based on the number of followers, so business owners are paying much more than they actually should [8]. The fake followers' percentage can go up to 78% of the total number of followers [9]. Fake followers make a user seemingly more popular than others [10], and this hurts the influencer's reputation [11].

Followers selling services have become a huge market in Indonesia [6]. The sellers can be easily found in local forums, websites, and Instagram. In order to do this research, these accounts were purchased. In an early observation, some of them were actively posting images, some were inactive, some were

spammers. Regardless of whether they are bots or human users, it is clear that they are controlled users because they can follow the target user when someone buys them. Based on the presented facts, in this research, there are three classes of fake followers, i.e. (1) active users, (2) inactive users, (3) spammers. It means that they are just filling up space, without ever interacting or buying advertised products.

Most existing research proposed a classifier model using Twitter as the platform, while only one report [12] was using Instagram. However, there is no existing research that classifies fake users into three classes (active, inactive, spammers), and no existing research uses location tags, hashtags, keywords and similarities in the media as features. The more detailed features and classifications in this study will help business owners in identifying the authenticity of their potential brand marketers.

The questions that will be answered in this research are (Q1) Are user's metadata and media data sufficient to create a machine learning model to identify fake followers? (Q2) What is the best classifier for fake users' classification? (Q3) What are the key differences between fake users and authentic users?. This research aims to identify different behaviors of authentic and fake users and propose a model to classify fake users on Instagram. This research only aims to classify fake users who have no value as followers. Classifying users of negative sentiments, such as identity impersonation attacks [13] and hatred [14], as some other studies have done is not considered. This research will use five supervised machine learning techniques for classification, i.e. Random forest, Neural Network, Logistic Regression, Naive Bayes, and J48 Decision Tree. These methods yielded the best accuracy in most research [15-17]. With an accuracy of up to 91.76% of the proposed model, the results of this study can help in eliminating fake users and make a healthier social media environment.

2. LITERATURE REVIEW

As stated earlier, fake users can also contain human users [6], and this expands the definition of a fake user. However, many studies consider fake users to be bot users only. The commonly used fake project dataset [18] was acquired from bot users market and CAPTCHA validation, so it contains only bot users. A list of fake users' classification is presented in Table 1. Most studies were using supervised, features-based detection method, and used Twitter as the platform. Supervised Machine Learning (ML) techniques to identify fake accounts in the fake project dataset were done in [15, 16, 19], with different features set. Unlike Twitter, Facebook is richer in terms of media-related features. Features such as tags, shares, comments, and likes (given and received) can be used for identification [20]. There is one report [12] that used Instagram for fake accounts classification. However, only metadata features were used, and the fake or authentic user decisions are based on human judgement, instead of getting the fake users from bot-selling markets. Instagram doesn't seem popular for research, despite its rapidly increasing popularity in influencer marketing [3].

Table 1. Recent research on fake accounts detection

#	Ref, Year	Platform/num fake	Method	Accuracy	Features used							
					M	MI	E	MT	MS	G	T	Total
1	[10], 2018	Twitter/9.7k	Supervised ML	97.75%	12	2	-	1	1	-	-	16
2	[19], 2018	Twitter/3.3k	SVM + NN	98%	7	-	-	1	-	-	8	16
3	[21], 2018	Twitter/1%	Reg., SVM	97.6%	-	-	-	-	-	1	-	1
4	[12], 2017	Instagram/7m	Random forest	94.4%	11	-	-	-	-	-	-	11
5	[20], 2017	Facebook/1.5k	Supervised ML	79%	1	16	-	-	-	-	-	17
6	[15], 2016	Twitter/3.3k	5 ML methods	99.5%	17	4	-	1	-	-	-	22
7	[22], 2016	Twitter/-	Profile-based analysis	-	10	-	-	-	-	-	-	10
8	[17], 2015	LinkedIn/15k	Cluster, Reg., SVM, RF	98.9%	5	-	-	-	-	-	-	6
9	[16], 2015	Twitter/1.9k	8 supervised ML	99.4%	21	13	-	1	-	-	-	35
10	[23], 2014	Real-time users	Clustering	99%	-	-	-	-	-	1	-	1

Notes:
Abbreviations: ML (machine learning), SVM (support vector machine), RF (random forest classifier)
Some features are derivatives of others, such as "has sent < 50 tweets" and "has never tweeted" derived from tweets count [10]. Below is the list of features code used in the table:

- M: Metadata, i.e. profile picture, user biography, link, number of followers, following, friends, etc.
- MI: Media info
- E: Engagement (Number of likes/comments divide by number of followers)
- MT: Media tags
- MS: Media similarity
- G: Relation analysis using graph
- T: Usage of themes on user page (only on Twitter), such as background, sidebar, text color

Similarities are important differentiators of fake and authentic users, such as friends [21], activities [23], names similarity [17], posts similarity [10]. Friends similarity can be used because fake accounts tend to have similar friends to each other [21]. However, it is not practical if the fake users are coming from different countries and have different set of friends. Activity similarity [23], on the other hand, is more suitable for detecting malicious actions. Twitter data pattern analysis [22] revealed that bot accounts were almost always created in batches within a very short interval. They also have similarities in screen names and post update frequencies. This kind of pattern similarity is also used by [17] to identify fake accounts, using clustering and supervised ML.

3. RESEARCH METHODOLOGY

This research starts with fake users and authentic users data collection. All private users were removed, because only user's metadata can be acquired from them, not media data. Available metadata on Instagram are username, full name, biography, link, profile picture, number of posts, following, followers. After data collection, these features will be extracted, and the correlation analysis will be carried out. After setting up the features to be used, machine learning algorithms will be used to classify the users. The complete research methodology is shown in Figure 1. There are four classes of users to be identified in this research, i.e. real users, and three classes of fake users (active, inactive, spammer). These classes are based on manual observation of their behaviors.

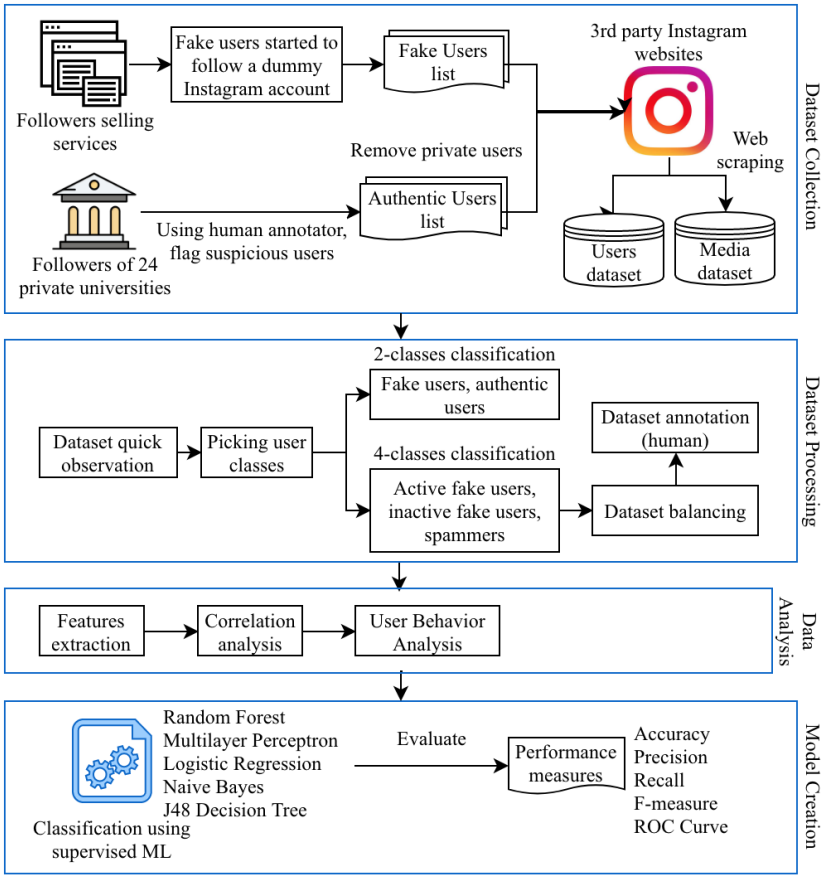

Figure 1. Research methodology

4. DATASET

This section will discuss about data collection and annotation, features extraction method, correlation analysis between features, and user behavior analysis based on average of each feature of each user class.

4.1. Data collection

The dataset was collected using web scraping from third-party Instagram websites, to capture their metadata and up to 12 latest media posts from each user. The collection process was executed from September 1st, 2019, until September 20th, 2019. The dataset contains authentic users and fake users, which were filtered using human annotators. The authentic users were taken from followers of 24 private university pages (8 Indonesian, 8 Malaysian, 8 Australian) on Instagram. To reduce the number of users, they are picked using proportional random sampling based on their source university. All private users were removed, which is a total of 31,335 out of 63,795 users (49.11%). The final number of public users used in this research was 32,460 users.

There are reasons for choosing private universities, i.e. (1) private universities tend to have lesser followers if compared to public universities, so the followers are most likely more authentic, (2) universities have fewer tendencies in buying followers, compared to influencers, (3) engagement rate of private universities are higher than public ones. Engagement rate can be used to measure the authenticity of followers [8, 24]. To prove the reason (1) and (3), the top 5 public and private universities in Malaysia are sampled and measured with FakeCheck.co. The average engagement rate (likes divide by followers) of private universities is 2.99%, compared to 2.43% of public universities. The average follower of private universities is 16,052, compared to 18,799 of public universities.

The fake users were collected by buying followers from Indonesia sellers, from various sources (from Instagram and Kaskus forum). In early observation, these followers satisfy the three types of fake users mentioned in section 1. Sometimes, bot users are quite obvious, usually they come with random names with random numbers in the username [17], no profile picture, no post, no biography. The data on the followers used in this research are from across the world. A dummy Instagram business account for the followers to follow was created to carry out this research. According to the analytics data in the dummy account, the top 5 countries of the fake followers are Indonesia (17%), India (13%), Turkey (9%), Pakistan (8%), Russia (7%). This indicates that this follower selling service is worldwide, and the results of this research are less likely biased towards specific countries.

There were two experiments to be conducted, i.e. with 2-classes classification (fake or authentic user), and with 4-classes. Thus, to balance the dataset, some authentic users are removed in the 4-classes classification. The removal is done by using K-Means to 10 clusters and proportional random sampling. From each cluster, 1,120 users will be picked randomly, as detailed in Table 2. This sampling is intended to preserve different user behavior. The 2-classes classification doesn't require annotation since the authentic users and fake users are coming from different sources. The 4-classes classification requires human annotation for the fake users' class. A simple web page was built to provide UI for the human annotator to annotate. The details of dataset numbers are provided in Table 3.

Table 2. Cluster result of authentic users (for 4-classes classification)

Cluster #	Number of users-Used in 2-lasses classification	%	Users picked (randomly)-Used in 4-classes classification
1	2,029	6%	1,120
2	2,712	8%	1,120
3	4,755	15%	1,120
4	3,032	9%	1,120
5	1,059	3%	1,059
6	422	1%	422
7	8,850	27%	1,120
8	2,351	7%	1,120
9	5,521	17%	1,120
10	1,729	5%	1,120
Total	32,460	100%	10,441

Table 3. Details of the dataset

#	Classification	Number of fake users	Authentic users	Total users	Media (fake)	Media (auth)
1	2-classes	32,869 fake users	32,460 users	65,329	376,357	460,923
2	4-classes	12,054 a, 10,549 i, 10,263 s	10,441 users	43,307	376,357	141,371

Notes:
- a = Active fake users, i = Inactive fake users, s = Spammers
- Media (fake) = Total media of all fake users in the dataset
- Media (auth) = Total media of all authentic users in the dataset

4.2. Dataset annotation

For the authentic users, to verify the authenticity, three human annotators were assigned to flag suspicious (or fake) users. Based on the majority voting, if two or more annotators flagged a user, the user will be removed. However, this relies on the human judgement. One more accurate approach is the CAPTCHA validation [18], which takes a lot of time to collect a large number of users. To classify the fake users to three classes, the same three human annotators were assigned. Prior to the annotation, some characteristics of each class of fake users were defined, i.e. (1) Active users, characterized by a high number of followers, have a lot of posts but usually short or no caption, or smiley-only caption. (2) Inactive users, which usually have no profile picture, and/or biography, and a low number of posts. (3) Spammers, which usually post a lot of images within a short interval. These characteristics serve only as hints, not mandatory conditions, for the annotators. The final decision is made based on majority of the voting. If there is no majority vote (6.31% occurrence), the annotator #1 decision is used as the priority.

A previous study [25] showed that there are 85% of users (from various Instagram users) who post up to 30 images. In our dataset, there are 56.4% of fake users who post up to 30 images, and 30.4% who post up to 5 images. For the spammers, upon early observations, there are generally three types of their posts, i.e. promotional or giveaway images, posts that attract new followers ("follow to follow" kind of), or any post that has high similarity with his/her other posts. Spammers can also be identified by post interval. In our media dataset, there are 22.15% of fake users who has post interval as short as 10 minutes between their own two latest posts. Overall, these facts are some additional hints for differentiating fake users to different classes.

4.3. Features extraction

The list of user features used for the classifier model is explained in Table 4. There are five different sources of the features, i.e. metadata, media info, engagement, media tag, media similarity.

Table 4. List of user features

#	Category	Var name	Feature name	Description
1	Metadata	*pos*	Num posts	Number of total posts that the user has ever posted.
2	(M)	*flg*	Num following	Number of following
3		flr	Num followers	Number of followers
4		*bl*	Biography length	Length (number of characters) of the user's biography
5		*pic*	Picture availability	Value 0 if the user has no profile picture, or 1 if has
6		*lin*	Link availability	Value 0 if the user has no external URL, or 1 if has
7	Media info	*cl*	Average caption length	The average number of character of captions in media
8	(MI)	*cz*	Caption zero	Percentage (0.0 to 1.0) of captions that has almost zero (<=3) length
9		*ni*	Non image percentage	Percentage (0.0 to 1.0) of non-image media. There are three types of media on an Instagram post, i.e. image, video, carousel
10	Engagement	*erl*	Engagement rate (Like)	Engagement rate (ER) is commonly defined as (num likes) divide by (num media) divide by (num followers)
11	(E)	*erc*	Engagement rate (Comm.)	Similar to ER like, but it is for comments
12	Media tags	*lt*	Location tag percentage	Percentage (0.0 to 1.0) of posts tagged with location
13	(MT)	*hc*	Average hashtag count	Average number of hashtags used in a post
14	Media	*pr*	Promotional keywords	Average use of promotional keywords in hashtag, i.e. {regrann, contest, repost, giveaway, mention, share, give away, quiz}
15	similarity (MS)	*fo*	Followers keywords	Average use of followers hunter keywords in hashtag, i.e. {follow, like, folback, follback, f4f}
16		*cs*	Cosine similarity	Average cosine similarity of between all pair of two posts a user has
17		*pi*	Post interval	Average interval between posts (in hours)

4.4. Correlation analysis

Correlation is a measure of association between variables, and high correlations among the independent variables produce an unusable model [26]. The bivariate correlation method will be used to produce the Pearson correlation values. The results are shown in

Table *5*. As shown in

Table *5*, there is no strong correlation (0.7 and up) between the variables. The highest values of correlation are *bl*/*lin* (0.47) and *erc*/*erl* (0.44), which are considered as moderate (0.40 to 0.69) and acceptable [27]. The correlation between *lin* (availability of link) and *bl* (biography length) shows that users with long biography information will most likely put a link. On Instagram, users can only put one link in the biography, and it's usually used to introduce his/her website or other social media. The *erc* (ER-comments) and *erl* (ER-likes) correlation shows that the number of likes correlates linearly with the number of comments.

Table 5. Correlation values (pearson correlation)

Var	*pos*	*flw*	*flg*	*bl*	*pic*	*lin*	*cl*	*cz*	*ni*	*erl*	*erc*	*lt*	*hc*	*pr*	*fo*	*cs*	*pi*
pos	-	0.14	0.06	0.16	0.05	0.17	0.19	-0.08	0.08	-0.03	-0.04	0.03	0.02	0.02	-0.01	-0.02	-0.09
flw	0.14	-	0.01	0.04	0.01	0.05	0.03	-0.02	0.03	-0.01	-0.01	0.01	0.01	0.00	0.00	-0.01	-0.01
flg	0.06	0.01	-	0.01	-0.13	-0.03	-0.06	0.17	-0.08	0.02	-0.02	-0.12	-0.04	-0.05	0.02	0.23	-0.09
bl	0.16	0.04	0.01	-	0.17	**0.47**	**0.35**	-0.27	0.14	-0.04	-0.06	0.22	0.16	-0.03	0.02	-0.14	-0.11
pic	0.05	0.01	-0.13	0.17	-	0.12	0.12	0.06	0.13	-0.02	-0.02	0.13	0.06	0.02	0.00	-0.27	0.08
lin	0.17	0.05	-0.03	**0.47**	0.12	-	0.30	-0.24	0.14	-0.05	-0.07	0.20	0.09	-0.04	-0.01	-0.10	-0.10
cl	0.19	0.03	-0.06	**0.35**	0.12	0.30	-	**-0.35**	0.11	-0.04	-0.05	0.08	0.19	0.21	0.06	-0.09	-0.11
cz	-0.08	-0.02	0.17	-0.27	0.06	-0.24	**-0.35**	-	-0.13	0.08	0.09	-0.21	-0.22	-0.06	-0.05	0.32	0.06
ni	0.08	0.03	-0.08	0.14	0.13	0.14	0.11	-0.13	-	-0.02	-0.03	0.22	0.05	-0.02	-0.01	-0.24	0.00
erl	-0.03	-0.01	0.02	-0.04	-0.02	-0.05	-0.04	0.08	-0.02	-	**0.44**	-0.02	0.02	-0.01	0.03	-0.03	0.00
erc	-0.04	-0.01	-0.02	-0.06	-0.02	-0.07	-0.05	0.09	-0.03	**0.44**	-	-0.02	0.03	0.02	0.03	-0.07	0.02
lt	0.03	0.01	-0.12	0.22	0.13	0.20	0.08	-0.21	0.22	-0.02	-0.02	-	0.13	-0.06	0.00	-0.27	0.07
hc	0.02	0.01	-0.04	0.16	0.06	0.09	0.19	-0.22	0.05	0.02	0.03	0.13	-	0.12	**0.34**	-0.15	0.02
pr	0.02	0.00	-0.05	-0.03	0.02	-0.04	0.21	-0.06	-0.02	-0.01	0.02	-0.06	0.12	-	0.08	-0.05	-0.02
fo	-0.01	0.00	0.02	0.02	0.00	-0.01	0.06	-0.05	-0.01	0.03	0.03	0.00	**0.34**	0.08	-	-0.02	-0.01
cs	-0.02	-0.01	0.23	-0.14	-0.27	-0.10	-0.09	0.32	-0.24	-0.03	-0.07	-0.27	-0.15	-0.05	-0.02	-	-0.14
pi	-0.09	-0.01	-0.09	-0.11	0.08	-0.10	-0.11	0.06	0.00	0.00	0.02	0.07	0.02	-0.02	-0.01	-0.14	-

Note: Five highest correlation absolute values are bolded

The three other variables with highest correlation, which is weak correlation (0.10 to 0.39), are *cl*/*cz* (-0.35), *cl*/*bl* (0.35), *fo*/*hc* (0.34). The *cl* (caption length) and *cz* (caption zero) are predictable, but the *cz* is used to strengthen *cl*. Fake users will usually post media without almost zero caption (<=3 characters, either no caption or smiley only caption). In the dataset, the percentage of media posted by fake users with almost zero caption is 32.8% (123,569 of 376,357), whereas for authentic users it's only 13.8% (63,470 of 460,923). The correlation between *cl* (caption length) and *bl* (biography length) indicates that users with long captions in media will probably put longer biography as well. The correlation between *fo* (follower keywords in hashtag) and *hc* (hashtag count) is predictable since *fo* is a subset of *hc*.

4.5. User behavior analysis

In this section, analysis of the different behaviors between authentic and fake users, and between the three classes of fake users is detailed. Statistics analysis results are shown in Table 6 and Table 7.

Table 6. Descriptive statistics of features

User	Stat	*pos*	*flw*	*flg*	*bl*	*pic*	*lin*	*cl*	*cz*	*ni*	*erl*	*erc*	*lt*	*hc*	*pr*	*fo*	*cs*	*pi*
Authentic	Avg	192	1661	1278	74	0.99	0.4	152	0.2	0.2	17.9	1.3	0.3	0.6	0.010	0.033	0.19	645
(32,460)	Stdev	667	30032	1692	65	0.11	0.5	208	0.3	0.3	33.9	3.0	0.3	1.1	0.068	0.387	0.26	1092
Fake	Avg	162	711	3331	41	0.92	0.1	121	0.3	0.2	20.4	1.0	0.1	0.4	0.056	0.073	0.41	350
(32,866)	Stdev	775	6742	2904	59	0.28	0.3	222	0.4	0.2	167.3	7.6	0.2	1.2	0.303	0.622	0.39	745
- Active	Avg	186	830	3142	51	0.99	0.1	118	0.3	0.2	9.9	0.4	0.2	0.3	0.014	0.009	0.22	399
(12,054)	Stdev	373	2757	2720	61	0.10	0.4	174	0.3	0.3	26.1	0.8	0.3	0.6	0.046	0.028	0.31	537
- Inactive	Avg	2	216	3817	12	0.77	0.0	12	0.3	0.1	38.2	1.9	0.0	0.1	0.001	0.000	0.64	316
(10,549)	Stdev	3	566	3180	36	0.42	0.2	58	0.4	0.2	290.1	12.7	0.2	0.6	0.010	0.008	0.46	1003
- Spammer	Avg	297	1081	3053	59	0.98	0.2	237	0.3	0.2	14.3	0.7	0.1	0.8	0.161	0.222	0.38	328
(10,263)	Stdev	1309	11658	2751	65	0.13	0.4	305	0.3	0.2	43.0	4.4	0.2	1.9	0.524	1.098	0.26	631

Below is the analysis of the variables:

- Metadata (*pos*, *bl*, *pic*, *lin*): Fake users have almost the same number of posts as authentic users. However, the spammers have the significantly biggest number of posts. Authentic users have the longest biography text, and 45% of authentic users provide a link. Only 77% of inactive users have a profile picture, whereas other user categories almost always have a profile picture.
- Follow info (*flw*, *flg*): Authentic users have the highest *followers* count, but lowest *following* count. In contrast, fake users have a lower *followers* count, but a higher *following* count if compared to the authentic users. This indicates that fake users like to follow others to increase their presence.
- Engagement (*erl*, *erc*): Fake users will receive more likes if compared to authentic users. However, in terms of comments, authentic users receive more. This indicates that receiving comments is harder, so the authentic users win in that case. In terms of likes, fake users generally have other fake users following them [21], so they can receive automated likes.
- Media info (*cl*, *cz*, *ni*): Authentic users have longer captions, and have less zero caption if compared to fake users. They also have higher video/carousel posts since it is easier for fake users to post a single image.

- Media tags (*lt*, *hc*): Authentic users have higher use of location tags and hashtags if compared to fake users. However, spammers use more hashtags if compared to authentic users, to attract users.
- Media similarity (*pr*, *fo*, *cs*, *pi*): Authentic users have less *cs* value. It means their posts are mostly different from their previous posts, unlike fake users. Spammers have the highest *pr* and *fo*. Authentic users will post media on average every 644 hours, which is almost twice slower than fake users (every 350 hours).

Table 7. Percentage of difference of average between authentic and fake users, and between active fake users and other fake users

User	*pos*	*flw*	*flg*	*bl*	*pic*	*lin*	*cl*	*cz*	*ni*	*erl*	*erc*	*lt*	*hc*	*pr*	*fo*	*cs*	*pi*
Authentic User as Reference																	
Fake	-16	-57	161	-45	-7	-75	-20	50	0	14	-23	-67	-33	460	121	116	-46
Active User as Reference																	
- Inactive	-99	-74	21	-76	-22	-100	-90	0	-50	286	375	-100	-67	-93	-100	191	-21
- Spammer	60	30	-3	16	-1	100	101	0	0	44	75	-50	167	1050	2367	73	-18

5. CLASSIFICATION

In this step, five supervised machine learning algorithms are used for the classification tasks, with the 17 mentioned features. The classification will be divided into 2-classes and 4-classes classification. The outcomes of each classification are the standard performance measures [28], i.e. accuracy, precision, recall, F-measure, ROC curve. Commonly, a smaller number of classes will lead to better accuracy, as shown in comparison in [29]. The result of classification algorithms with 10-fold cross-validation is shown in Table 8.

Table 8. Classification result (metrics in %)

#	Algorithm	2-Classes Classification					4-Classes Classification				
		Acc	Prec	Recall	F-mea.	ROC	Acc	Prec	Recall	F-mea.	ROC
1	Random Forest	90.09	90.7	90.1	90.1	96.3	91.76	91.7	91.8	91.7	99.1
2	Multilayer Perceptron	81.73	81.8	81.7	81.7	89.6	73.75	73.8	73.7	73.5	91.1
3	Logistic Regression	80.94	81	80.9	80.9	88	68.54	68.1	68.5	68.1	89.5
4	Naive Bayes	73.12	75.9	73.1	72.4	83.2	54.22	60.6	54.2	49.3	81.9
5	J48 Decision Tree	88.34	88.6	88.3	88.3	92.5	88.28	88.2	88.3	88.2	96

As shown in Table 8, Random Forest consistently outperforms other algorithms. Interestingly, while other algorithms struggle in the 4-classes classification, Random Forest can perform even better than the 2-classes counterpart. Also, the features importance can be acquired from the Random Forest calculation, as shown in Table 9. The features importance result is important to understand what the highest predictors are to differentiate fake users from authentic users. An additional insight can be acquired by combining the result of Table 9 and Table 7. For the 2-classes classification, the values of *flg* and *lin* are among the top five important features, as well as the top five behavioral differentiators between authentic and fake users.

Table 9. Features importance (order by importance in 2-classes)

Variable	*pos*	*flw*	*lin*	*flg*	*bl*	*cl*	*cz*	*ni*	*erc*	*erl*	*lt*	*hc*	*pic*	*pi*	*cs*	*fo*	*pr*
2-Class	0.42	0.39	0.37	0.36	0.35	0.33	0.32	0.31	0.3	0.3	0.27	0.27	0.25	0.25	0.25	0.19	0.18
4-Class	0.46	0.42	0.3	0.38	0.38	0.38	0.32	0.31	0.32	0.33	0.26	0.29	0.25	0.29	0.3	0.29	0.25

In the 4-classes classification, *pos*, *flw*, *bl*, *flg* are the four highest predictors. In the 2-classes classification, *pos*, *flw*, *lin*, *flg*, *bl* are the five highest predictors. These are all metadata values which are easy to acquire, even for private users. In the next experiment, classification using metadata-only variables will be carried out, to see if it is possible to classify fake and authentic users using only metadata. Table 10 shows the result of 2-class and 4-class classifications using six metadata variables (*pos*, *flg*, flr, *bl*, *pic*, *lin*). The classification with metadata as shown in Table 10 produced a less accurate result if compared to the classification with all features included as shown in Table 8. The highest accuracy that can be achieved in Table 10 was 79.66% (for 2-classes) and 59.14% (for 4-classes). The lower accuracy result of the 4-classes classification is because the differentiation of fake user types highly relies on media data.

Table 10. Classification result with metadata variables only

#	Algorithm	2-Classes Classification					4-Classes Classification				
		Acc	Prec	Recall	F-mea.	ROC	Acc	Prec	Recall	F-mea.	ROC
1	Random Forest	78.75	78.9	78.8	78.7	**87**	57.75	56.6	57.8	57	**83**
2	Multilayer Perceptron	73.86	73.9	73.9	73.9	82.6	46.48	45.3	46.5	43.4	72.3
3	Logistic Regression	73.44	73.7	73.4	73.4	82	55.23	53.7	55.2	51	80.6
4	Naive Bayes	66.62	72.2	66.6	64.3	79.8	43.84	45.8	43.8	32.9	73.5
5	**J48 Decision Tree**	**79.66**	**80**	**79.7**	**79.6**	86.5	**59.14**	**57.1**	**59.1**	**57.2**	81.4

6. DISCUSSION

The dataset analysis result provides some important insights, such as users with long biography will more likely put a link as well, and the number of likes is linear with the number of comments. The biography and link itself, are two of the five highest differentiators of authentic and fake users, with the importance of 0.35 and 0.37 (for 2-classes classification), respectively. Fake users have less biography length (average 41.3 characters) if compared to authentic users (average 73.9 characters), and less likely to put a link (12% vs. 45%). Another important differentiator is, fake users have a higher number of *following* if compared to authentic users, but in contrast, fake users have a lower number of *followers* if compared to authentic users. The provided 17 features for classification are proved to be sufficient to differentiate authentic and fake users. The highest classification accuracies for 2-classes and 4-classes are 90.09% and 91.76%, respectively, both using Random Forest. Furthermore, the features importance result indicates that the most important predictors in the classification are all metadata features, i.e. number of posts, number of followers, link availability, number of following, and biography length.

The high importance of metadata features leads to another experiment, which is classification using metadata-only features. This classification produced the highest accuracy of 79.66%. The advantage of the classification using metadata-only is, the metadata are easily acquirable, even for private users. This is a huge advantage since in the raw dataset, 49.11% of the users are private users. Thus, even though using the full features can produce a better classification accuracy, the classification with metadata-only is still acceptable given the fact that a lot of Instagram users are private users. Furthermore, collecting metadata from many users is much faster than collecting both metadata and media data.

7. CONCLUSION

The outcome of this research is important for business owners who look forward to finding influencers for brand endorsement. Unlike Twitter, which is used by most research, Instagram has richer features in media sharing, and it is proven to be the most used platform for brand marketing. Possible improvement for this research is the inclusion of text analysis in both caption and comments, relation graph analysis, and image analysis. Some captions of the fake users are irrelevant, such as "follow me", and some comments given to them are also coming from fake users. By doing this, the fake users can avoid being banned by Instagram by keeping good engagement levels. Image analysis is also useful since many spammers post text-based images. Graph analysis is also helpful for the classification since some of the fake users are in the same circle of relation.

REFERENCES

[1] Statista, "Most famous social network sites worldwide as of July 2018, ranked by number of active users (in millions)," 2018. [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. [Accessed 3 October 2018].
[2] Lyfe Marketing, "The Best Social Media Platforms for Social Media Marketing in 2018," Lyfe Marketing, 2018.
[3] Rakuten Marketing, "Influencer Marketting Global Survey," Rakuten Marketing, San Mateo, California, 2019.
[4] Y. Y. A. Talib and R. M. Saat, "Social proof in social media shopping: An experimental design research," in *SHS Web of Conferences*, vol. 34, 02005, 2017.
[5] M. Code, "Instagram, Social Media, and the "Like": Exploring Virtual Identity's Role in 21st Century Students' New Socialization Experience," Faculty of Education, Brock University, St. Catharines, Ontario, 2015.
[6] J. Lindquist, "Illicit Economies of the Internet: Click Farming in Indonesia and Beyond," Made In China: To The Soil, 2018.
[7] S. Rodriguez, "Instagram Could Delete Up To 10 Million Accounts As It Cracks Down On Spam," *International Business Times*, 12 November 2014. [Online]. Available: https://www.ibtimes.com/instagram-could-delete-10-million-accounts-it-cracks-down-spam-1749914. [Accessed 4 September 2019].
[8] E. G. Ellis, "Fighting Instagram's $1.3 Billion Problem—Fake Followers," Wired, 9 October 2019. [Online]. Available: https://www.wired.com/story/instagram-fake-followers/. [Accessed 14 September 2019].

[9] J. Neff, "Study of Influencer Spenders Finds Big Names, Lots of Fake Followers," 23 April 2018. [Online]. Available: https://adage.com/article/digital/study-influencer-spenders-finds-big-names-fake-followers/313223. [Accessed 14 September 2019].
[10] P. G. Efthimion, S. Payne and N. Proferes, "Supervised Machine Learning Bot Detection Techniques to Identify Social Twitter Bots," SMU Data Science Review, vol. 1, no. 2, 2018.
[11] C. Lieber, "The Dirty Business of Buying Instagram Followers," 11 September 2014. [Online]. Available: https://www.vox.com/2014/9/11/7577585/buy-instagram-followers-bloggers. [Accessed 14 September 2019].
[12] J. V. Scurrell and T. Grossenbacher, "Identifying a Large Number of Fake Followers on Instagram: A Statistical Learning Approach," DIWA-Capital, 2017.
[13] E. Villar-Rodriguez, J. D. Ser and S. Salcedo-Sanz, "On a Machine Learning Approach for the Detection of Impersonation Attacks in Social Networks," in *Intelligent Distributed Computing VIII*, Cham, Springer, pp. 259-268, 2015. Doi: 10.1007/978-3-319-10422-5_28.
[14] T. Davidson, D. Warmsley, M. Macy and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, Association for the Advancement of Artificial Intelligence, pp. 512-515, 2017.
[15] A. E. Azab, A. M. Idrees, M. A. Mahmoud and H. Hefny, "Fake Account Detection in Twitter Based on Minimum Weighted Feature set," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 10, no. 1, pp. 13-18, 2016.
[16] S. Crescia, R. D. Pietrob, M. Petrocchia, A. Spognardia and M. Tesconia, "Fame for sale: efficient detection of fake Twitter followers," *Decision Support Systems*, vol. 80, pp. 56-71, 2015.
[17] C. Xiao, D. M. Freeman and T. Hwa, "Detecting Clusters of Fake Accounts in Online Social Networks," in *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*, ACM, pp. 91-101, 2015.
[18] "The Fake project," [Online]. Available: http://wafi.iit.cnr.it/theFakeProject/. [Accessed 30 October 2015].
[19] S. Khaled, H. M. O. Mokhtar and N. El-Tazi, "Detecting Fake Accounts on Social Media," in *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, 2018.
[20] A. Gupta and R. Kaushal, "Towards Detecting Fake User Accounts in Facebook," 2017.
[21] M. Mohammadrezaei, M. E. Shiri and A. M. Rahmani, "Identifying Fake Accounts on Social Networks Based on Graph Analysis and Classification Algorithms," *Security and Communication Networks*, 2018. doi: 10.1155/2018/5923156.
[22] S. Gurajala, J. S. White, B. Hudson, B. R. Voter and J. N. Matthews, "Profile characteristics of fake Twitter accounts," *Big Data & Society*, pp. 1-13, 2016. doi: 10.1177/2053951716674236.
[23] Q. Cao, X. Yang, J. Yu and C. Palow, "Uncovering Large Groups of Active Malicious Accounts in Online Social Networks," *Proceedings of the ACM Conference on Computer and Communications Security*, pp. 477-488, Scottsdale, 2014. doi: 10.1145/2660267.2660269.
[24] C. Forsey, "Why You Shouldn't Buy Instagram Followers (& What to Do Instead)," HubSpot, 11 July 2019. [Online]. Available: https://blog.hubspot.com/marketing/buy-instagram-followers. [Accessed 14 September 2019].
[25] C. S. Araujo, L. P. D. Corrêa, A. P. C. d. Silva, R. O. Prates and W. M. Jr., "It is not just a picture: Revealing some user practices in Instagram," *in IEEE 9th Latin American Web Congress*, 2014.
[26] D. C. Montgomery, E. A. Peck and G. G. Vining, "Introduction to Linear Regression Analysis," John Wiley & Sons, Hoboken, New Jersey, United States: John Wiley & Sons, Inc., 2012.
[27] P. Schober, C. Boer and L. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation," *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763-1768, 2018.
[28] J. Davis and M. Goadrich, "The Relationship between PrecisionRecall and ROC Curves," in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, 2006.
[29] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini and A. Vakali, "Mean birds: Detecting aggression and bullying on twitter," in *Proceedings of the 2017 ACM on Web Science Conference, ACM*, pp. 13-22, 2017.

BIOGRAPHIES OF AUTHORS

Kristo Radion Purba, Kristo is currently a computer science PhD student at Taylor's University Malaysia, starting from 2018. His research interests are in artificial intelligence, machine learning, and social network influence maximization. Prior to joining Taylor's, he was an informatics lecturer at Petra Christian University, Indonesia for 4 years (2014-2018), and also a contracted programmer at EHS (Environment, Health and Safety) department at PT. HM. Sampoerna, Tbk, Indonesia (2013-2017). He is also an active mobile apps, games, websites developer since 2008 until now. Email: kristoradionpurba@sd.taylors.edu.my

Professor Dr David Asirvatham, Dr. David Asirvatham is currently the Head for the School of Computing and IT, Taylor's University. Prior to this, he was the Director for the Centre of Information Technology at University of Malaya. He has held numerous posts such the Associate Dean for Faculty of Information Technology (Multimedia University), Project Manager for the Multimedia and IT Infrastructure Development for a university campus (US$14 million), Finance Committee for Multimedia University, SAP Advisory Council, Consultant for e-University Project and many more. Dr. David completed his Ph.D. from Multimedia University, M.Sc. (Digital System) from Brunel University (U.K.), and B.Sc. (Hons) Ed. and Post-Graduate Diploma in Computer Science from University of Malaya. He has been lecturing as well as managing ICT projects for the past 25 years. His area of expertise will include Neural Network, E-Learning, ICT Project Management, Multimedia Content Development and recently he has done some work on Big Data analytics. Email: david.asirvatham@taylors.edu.my

Associate Professor Dr Raja Kumar Murugesan, Dr Raja Kumar Murugesan is an Associate Professor of Computer Science, and Head of Research for the Faculty of Innovation and Technology at Taylor's University, Malaysia. He has a PhD in Advanced Computer Networks from the Universiti Sains Malaysia and has over 28 years' experience as an educator. His research interests include IPv6, and Future Internet, Internet Governance, Computer Networks, Network Security, IoT, Blockchain, Machine Learning, and Affective Computing. He is a member of the IEEE and IEEE Communications Society, Internet Society (ISOC), and associated with the IPv6 Forum, Asia Pacific Advanced Network Group (APAN), Internet2, and Malaysia Network Operator Group (MyNOG) member's community. Email: rajakumar.murugesan@taylors.edu.my