A dataset for the detection of fake profiles on social networking services.

1st Samuel Delgado Muñoz
Systems Engineering Department
El Bosque University
Bogotá, Colombia
sdelgadom@unbosque.edu.co

2nd Edward Paul Guillén Pinto
Systems Engineering Department
El Bosque University
Bogotá, Colombia
eguillenp@unbosque.edu.co

***Abstract*—The use of multiple social media platforms is a common practice on more than two-third of all Internet users, according to OurWorld In Data. From this perspective, the verification of a real profile is a matter of growing interest, because false virtual identity could trigger problems such as spoofing, bots, grooming, sextortion, just to name a few. This paper presents a method to detect fake profiles on social media platforms by deploying some machine learning detection methods over a novel dataset. The dataset was designed with 17 metadata features from real and fake profiles and it was tested on Instagram profiles. After deploying different machine learning algorithms, the obtained detection rate was near to 96% with good false positive rates.**

***Index Terms*—Dataset, profile detection, unsupervised lerning, social networks, fake profiles, machine learning**

I. Introduction

Online impersonating or identity theft has become a bigger problem with each year that passes. According to [1] impersonating and phishing are becoming uprising threats because of the evolution of technology. This evolution is helping these kind of attacks to become more and sofisticated and thats why security technologies have to evolve too. Also, the consecuences of this kind of online attacks may keep becoming worse as their scope grows, embracing the stealing of personal information and money.

According to the proposal of [2], this type of attack mostly represents a challenge for authentication methods in different types of networks and also for the early detection of cases of impersonation and phishing. This can be achieved through the implementation of different technologies such as PHY security in IoT systems. All this taking into account the factors of efficiency and robustness of authentication.

Impersonation, phishing and other kind of problems can be found in social networks too and they are becoming the most used tool for this malicious purposes because of the social networks increasing usage. As said in [3], social networks such as Facebook, Twitter, LinkedIn and tumblr are some of the most used social networks having millions of connections in just one day.

Millions of people using social networks are mostly unaware of the dangers and security risks that exists in this types of communications. These risks include privacy risks, fake profiles, sexual harassment among others.

Personal information stealing and impersonation are strongly related problems as one can derivate from the other. In [3], expose a hypothetical of creation of a fake Facebook profile and the privacy consequences it may have is included. Exposed step by step, the attacker created a profile with detailed false information and tried to establish contact with the victim by sending a friend request. In this case, if the victim agrees, part of the information is automatically exposed.

Additionally if the attacker tries to dig deeper, more sensitive information can be exposed to an unwanted third party. All the information extracted using this technique, such as lists of friends, places of work and study, can be used for future and wider attacks or simply as a collection of valuable information.

Fake profiles are not always managed by a person, these profiles can be bots that mimic human behavior and are in charge of accumulating information from different users. In [3] it is mentioned that these bots are also often used to cause massive spam, manipulation of social network statistics and sudden increases in server traffic. additionally, friend lists are often used to expand and multiply the impact of the consequences of these bots.

Another bot related attack that can be made using any social networks fake accounts it's called Sybil attack. According to [4] this kind of attack seeks for establishing a huge and solid fake accounts network to later attempt to manipulate the system in ways it is not meant to be manipulated. In fact, this is contrary to the most common rule on social media wich dictates that one user can have one account only. After establishing these fake accounts into the network by sending an enormous quantity of friend requests or follow request, these fake accounts become more a more trustfull cause they have been building relevance in the selected platform. Additionally as stated in [4], this accounts can be used to manipulate poll results and for spreading false information massively.

As mentioned before, fake social networks accounts and impersonation attacks can have a lot of consequences when speaking about security but these are not the only consequences they may have. In another completely different

Project funded by El Bosque University

978-1-7281-7624-6/20/$31.00 ©2020 IEEE
DOI 10.1109/CSCI51800.2020.00046

context, fake accounts and impersonation can be used to psychologicaly hurt someone just as stated in [5]. Even when this scenario is present mostly in teenagers, fake accounts and impersonation has a big role in cyberbullying and therefore in teenagers psychological health and mental stability.

Taking into account all the social networks information exposed before, the focus of this paper is to extract metrics that allow the recognition and differentiation of legitimate profiles and fake profiles on an specific social network: Instagram. According to Statistical [6] in June 2018 Instagram had over a thousand million of active users making it one of the most used social networks. Photographs and videos are shared in both personal and advertising spheres. In the personal section, this social network is widely used to share passions, trips, experiences, and to connect with other people. Due to its great popularity, and just as the social networks mentioned before, some malicious users create false profiles seeking to deceive users by posing as people who they are not.

Currently, the identification of these fake profiles is limited to manual procedures. It is a tedious process and may not be sufficient due to the quality with which they can duplicate existing profiles or make compilations to create profiles bots, spammers, phishers, impersonations, or fake accounts [4]. At present, and following the main topic of [2] and [3] there are many research approaches to identify fake profiles and other kind of threats on social net-works such as Facebook or Twitter. The ease with which these platforms offer application programming interfaces (API) makes it easy to obtain up-to-date, real-time profile data.

Instagram is a social network with stricter privacy policies, where there are proven users with very restricted visibility. The objective of this research, as mentioned before is to gather metrics and data making use of the information of users of public and private access.

To carry out this project, a set of true and false profiles is required to create a binary decision dataset. For this, there were two main strategies: the use of simulated profiles that could emphasize too many cer-tain features, which could bring possible problems of simulated data. The second strategy is to use web sources that have exposed and verified false profiles. This was the one used and several web sources were found where these profiles were identified and validated manually assisting research with verified profiles.

II. Related Work

Much effort has been put into fake profiles and spoofing analysis on different social networks due to the increasing the threat it poses to users. There have been many different approaches over the years making use of different and newer technologies as they and social media evolve as well.

To illustrate how there has been a lot of approaches when speaking about social networks data analysis, we can take a look at the made by [7]. In this case, this paper exposes the different ways the data stored in social networks can be accessed and treated. In this paper the authors explain that treating this information is always a challenge because social networks data can get massive. To make it easier, they state that social networks information can be divided in two groups being structured and unstrictured data. The example they expose is that real time events are structured data and things such as retweets and reactions are unstructured data. They also mention Artificial Intelligence approaches that make of extracting statistics, methods based in content analytics, text mining among others. In conclusion, they found that every approach deserves to be further developed because eachone offers a different point of view and a different way to be implemented.

An interesting approach to the problem of detecting fake accounts in social networks is the one proposed by [4]. In this article they take a point of view based on the victims of this fake accounts to construct a detection system. In this case, they presented Íntegro, as a software based on a raking scheme graph based algorithm but also, making use of some final user activites. They state that this process is completely transparent for the user and that Íntegro works on social networks that only aproves bidirectional frienships. Specifically speaking, they make use of users activities and information that is cheap to find and extract. After taking this information, they use it to train a victim classificator that allows Íntegro to find potential victims and start the fake account identification from that start point. As a conclusion, they claim that Íntegro proved to be an efficient way to find fake accounts because of the different aproach of using the victims accounts as a resource to find the fake ones.

The work made in [4] exposes how they tried to predict when and how sybil attacks where going to be made. In this specific case, they proposed the development of a deep-learning regression model that allowes them to predict sybil attacks. They certainly got a focus in two different kinds of sybil attacks such as targeted attacks and automated attacks because they found this are the most hard to fight sybil attacks. They also stated that this model is totally focused on detecting malicious users in social networks to prevent sybil attacks. They selected Twitter as their target social network and extracted specific features from tweets and accounts to use them as a dataset to train their prediction neural network model. This model is based in the analysis and extraction of tweets content and each account information and actions done. In the end, the work done in this paper reached an 86% accuracy for predicting sybil-attacks, but they also stated that as this kind of attacks keep evolving, the proposal of countermeasures have to keep raising and evolving too.

As stated before in this paper, problems such as sybil attacks, social phishing and impersonation are rising threats for users of online social networks. These problems have a common source an that is the fake accounts that anyone can make by following a few steps and using a completely new e-mail or even a phone number. The work done in [8] exposes how this problems can be aborded by using artificial intelligence to enhance social networks security. Even though in this case they don't specifically speak of a way to solve the problems mentioned before, they propose a way

to treat information in social networks. This way is through crowdsourcing, a way to understand problems differently.

Twitter is a social network where a lot of these problems concentrate, and a lot of effort has been put to control fake and spam accounts. The proposal made in [9] has a different method approach. In this case, they concentrated their effort in the clustering of Spam accounts by organizing these groups by taking into account similar characteristics. Spammer accounts can also be fake automated accounts encharged of posting fake information and follow a lot of accounts to increase their reach. To gather the needed information they used a crawler that seeked for spam trigering words in a lot of real tweets during 2 months. Then, they extracted 15 features using Principal Component Analysis and worked with them to find the clusters of accounts that share the same kind of spam tweet sematic by using the k-means algorithm and then trained 3 different classificators with this data. In the end of their experiment, they reached a 96% accuracy by using a Ramdon Forest Classifier.

A pretty recent work regarding various amount of social networks and social media problems is the one performed in [10] where an automated model for forensic social network investigations was proposed. This paper is not really focused on the identification of fake accounts but in a more general view of a complete profile deep analysis. The proposed model in this article is semi-automated therefore, some decissions are still made bu human beings. In this speciffic case and due to legal conditions, most of the data extraction, data mining and data normalization cannot be used so they had to propose a model where automated processes weren't everything. In this kind of social network profiles analysis, early extrations and analyzes are performed by making use of software tools. The extraction phase is carried away by using a parser that takes incident information as an input and makes a first filter. Then the analysis phase starts and different methods just as classiffiers can be applied to filter the information extracted before. In the end, this kind of complex system has a lot of different results and it also can give information related to the use of fake accounts or profiles if they're involved in cases that need a forensic investigation.

Another fake accounts and fake profiles detection fitted approach is the one shown in [11], where they analyze a way to detect malicious bots that in the end are fake accounts. They state that the problem of this bots as fake accounts, is that they can expand their scope a lot by posting fake news and by making fake relationships with real users. Focused only on twitter and using twitters URL features to extract information, this work stablished an approach where every account is considered on its own, and the trustworthiness of each tweet and follow action is analyzed too. After extracting the needed information from the URL's they use a Learning Automata algorithm to determine if an account is a bot or a user of the social network. In this case, it can be considered a way to use machine learning to detect malicious bots, but with a totally different method to extract information. Another example of bot detection carried away using twitter as the main social media is the one made in [12], where they mostly used one-class classificators made using machine learning supervised and unsupervised trainig using features selected from acounts tweet and retweets content as well as the follows.

TABLE I
RELATED WORK METHODS USED

#	Reference	Platform	Method	Accuracy
1	[7]	General	AI+ML+TA	-
2	[15]	Facebook - Tuenti	GA	92%
3	[4]	Twitter	DL	86%
4	[8]	General	AI+CS	-
5	[9]	Twitter	CL+ML	96%
6	[10]	General	Hybrid	-
7	[11]	Twitter	URL+LA	91-95%
8	[12]	Twitter	ML	90%
9	[13]	General	GA+ML	-
10	[14]	Facebook - Twitter - Youtube	CNN	95%

[AI] Artificial intelligence, [GA] Graph based algorithms, [TA] Test Analytics, [ML] Machine Learning, [DL] Deep Learning [CS] Crowdsourcing, [CL] Clustering, [Hybrid] System using AI and human interaction, [URL] URL information [LA] Learning Automata, [CNN] Convolutional Neural Network

Another work that focuses in the malicious bot detecting is the one made in [13], where three state of the art approaches used to detect malicious bots in different online social networks were presented. The first apporach mentioned is the graph based one. This apporach uses three main characteristics to identifie fake or bot accounts: the cut between sybil and honest region, the fast mixing nature of a bot that is a pretty often behaviour for a fake account and finally the social edges that can make a fake account recognizable. The second approach employs machine learning techniques. This work divides the machine learning apporach in three categories: supervised machine learning, unsupersvised machine learning and hybrid machine learning. Finally, they state that there are emergent apporaches that seek to enhance the ones mentiones before such as detection of coordinated attacks among others.

Getting back to fake account detection we have the work made in [14] where they proposed a learning model based on conventional and statistical machine learning algorithms. The difference here radicates in the fact that they didn't limit their work to detecting bots, but fake accounts managed by people too. They used a supervised machine learning model and used a Convolutional three layer Neural Network that allowed them to gather informative values with different and new inputs. Again, they didn't use a preset dataset, but extracted information of real profiles.This case was a complete success, this model produces AUC = 0.9547 and really is better than any other kind of conventional approach for detecting this fake and malicious accounts.

In (table) we can see a classification of the previous exposed papers sorted by date, selected platform, method and accuracy achieved if applicable.

III. METHODOLOGY

The implemented methodology was divided into four main stages, profile categorization, data collection, feature selection, ending with the classification with machine learning algorithms.

A. Categorization of a profile

The selected dataset features were extracted taken based on the metadata and multimedia information of the publications. These features are represented in table II.

Below are some features with process beyond extraction: nickname_contains_name (ncm) this feature is obtained by applying similarity measurement techniques between two character strings, in this case, the name and the username. use Jaro distance equation (1)

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \quad (1)$$

where:

$|s_i|$ is the length of the string s_i
m is the number of matching characters
t is half the number of transpositions

nickname_complexy this characteristic is obtained by going through the string of characters of the username and making a sum depending on the type of character (digit, uppercase, other). This sum is it is divided by the result of nickname_contains_name , afterwards, the result is divided again is divided by the number of characters in the string as shown in the equations 2 and 3

$$f(x) = \begin{cases} 3, \text{ if } x \text{ is in domain [0-9]} \\ 2, \text{ if } x \text{ is in domain [A-Z]} \\ 1, \text{ if } x \text{ is not domain [0-9,A-Z]} \end{cases} \quad (2)$$

$$nc = \frac{\left(\frac{\sum_{i=1}^{n} f(x_i)}{(n\text{cm})}\right)}{n} \quad (3)$$

persentage_completed_profile this feature is obtained from the complete percentage of various sections, these sections and their percentage value are listed below:

- Has at least one post (10 %)
- Has 10 or more posts (10 %)
- Has at least one follower (5 %)
- Follows at least one account (5 %)
- Has name (5 %)
- Has a description (15 %)
- Has a web page (5 %)
- Has a profile picture (20 %)
- A face appears in the profile photo (10 %)
- A language was detected (15 %)

photos_similarly this feature is extracted from the analysis of each publication, verifying if there are known faces and if any are found, the photo is marked as similar. In addition to each iteration, the faces of the publications are added to the list of acquaintances. In the end, the known publications are counted and divided over the total of publications. this process is represented in figure 1.

photos_similarity_internet this feature is given by an analysis of similar images on the internet, 6 random images are chosen from the total to perform the analysis. From the results collected, the URLs obtained were verified in case of belonging to Instagram or Facebook, they are skipped because they are social networks in which profiles are shared. A score of 1 is used for 3 URLs different from the origin to identify an image as copied. With the total score divided over the total of images analyzed, this process can be better visualized in the figure 2

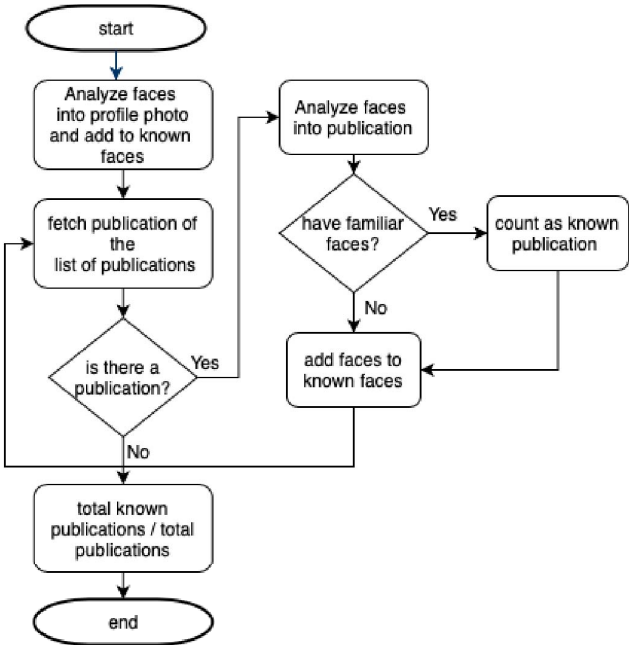

Fig. 1. process extract feature photos_similarly

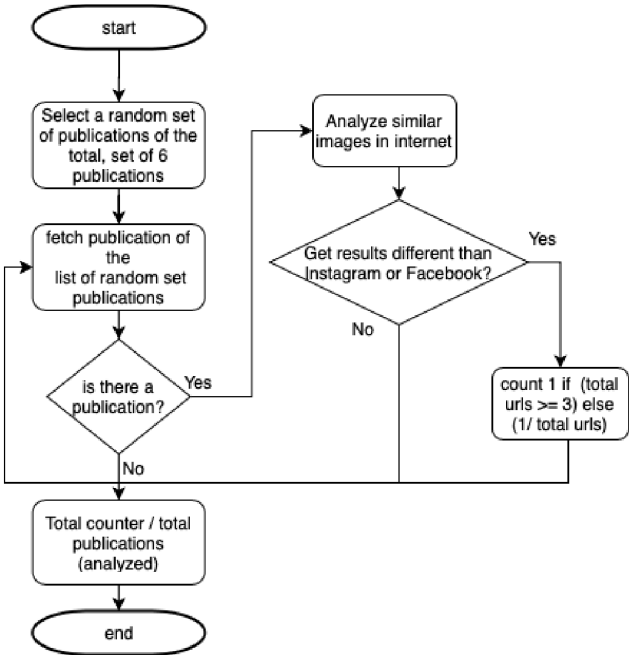

Fig. 2. process extract feature photos_similarity_internet

TABLE II
FEATURES OF USER

#	Diminutive	Name	Description	Type	Category
1	lc	location	Country and city of account (based in publications)	Nominal	MI
2	flws	followers	Followers of the user	Numeric	M
3	flwg	following	Following accounts to the user	Numeric	M
4	tp	total_publications	Total of publications of the account	Numeric	M
5	fp	first_publication_date	First publications date	Date	MI
6	lp	last_publication_date	Last publications date	Date	MI
7	ppw	publications_per_week	It's a average of total publications divided by # weeks between fp and lp	Float	MI
8	tp	tagging_publications	It's an array of the best labeling in publications analysis	Nominal	MI
9	lg	lang	Language of meta info only (en) and (es)	Nominal	M
10	nc	nickname_complexy	Nickname complexity	Numeric	M
11	ncm	nickname_contains_name	Nickname contains name	Numeric	M
12	pcp	persentage_completed_profile	Percentage completed profile	Numeric	M MI
13	uhp	user_has_photo	True if user has profile photo else False	Boolean	M
14	pc	private_account	True if account its private else False	Boolean	M
15	ppp	profile_photo_person	True if profile photo has a least person else False	Boolean	M
16	ps	photos_similarity	metric between 0 and 1 with similar images - based on posts	Numeric	M MI
17	psi	photos_similarity_internet	metric between 0 and 1 with similar images found in the internet	Numeric	MI

[M] Metadata
[MI] Media Information

B. Data collection

Web scrapping techniques were used for data extraction on thrid party sited to Instagram. The technologies used were Python and Selenium for the web scraping section. For the analysis of the publications we used the Google Vision API. The analyzed real profiles were selected from the followers of official pages of the programming league and profiles close to them were verified manually. The fake profiles were extracted from verified forums and publicly published profiles as fake, these profiles were also manually validated. From this analysis, 936 true profiles and 150 false profiles were extracted.

C. Feature selection

Data transformation was carried out by converting nominal data to numeric in the case of Boolean columns. then a standard scaling was used eliminating the mean and scaling to the variance of the unit $z = (x - u)/s$ where u is the mean of the training sample and s is the standard deviation of the training samples.

for this selection some features are extracted that for this iteration do not have much value (fp, lp, tp) since they are nominal and are used for other characteristics. With the remaining characteristics, a correlation analysis was carried out, which is the measure of association of the variables, the bivariate correlation method (Pearson) be used and the results. The results are shown in table III

In this analysis it can be seen that there is only a strong correlation between two characteristics (pcp, uhp), the other characteristics are not strongly correlated, which can make models unusable [16].

D. Classification

To analyze the dataset results, several tests were car- ried out with the following classification algorithms for each were taken into account Accuracy, Precision true, Precision false. A random sample of 20 percent of the dataset data was used to test the model and the remaining 80 percent as training data.

1) Decision Tree: They are a method of machine learning, it consists of creating a model that, based on a series of rules, predicts the value of a variable. This method is good, but you have to be careful with the generation of very complex trees that do not generalize the data well. This is called overfitting, there are pruning mechanisms to avoid this problem [17].

In the implementation there were no problems of overfitting, the results of this test can be seen in the table IV.

2) Logistic Regression: Logistic regression is a good test to verify if problems can be solved with a statistical solution before implementing machine learning. The other non-linear algorithms used in the tests have advantages and disadvantages with linear regression how do you explain it [18] and the results of these implementations can be seen in the table IV, some algorithms offer better results. binary logistic regression was used in this implementation

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + b_2 x_2 + \dots \tag{4}$$

where:

- p is the probability (risk) event occurring
- x are the independent variables
- b are the coefficients associated with each variable

3) Random Forest: Is a perturbation and combination algorithm, that creates a set of random classifiers, introducing randomness as a key factor [19]. The goal of this randomness is to decrease the variance of the estimator. Random forests achieve reduced variation by combining multiple trees, sometimes at the cost of slightly increasing bias.

4) Multi-layer Perceptron: It is an algorithm that uses the function shown below 5 as a basis for learning, it differs from a logistic regression since there may be one or more non-linear layers between the input and output layers.

$$f(\cdot) : R^m \to R^o \tag{5}$$

TABLE III
CORRELATION VALUES (FEATURES)

	lc	flws	flwg	tp	ppw	lg	nc	ncm	pcp	uhp	pc	ppp	ps	psi
lc	1	-0,09	-0,07	-0,09	-0,15	0,09	-0,01	-0,04	-0,18	-0,11	0,33	-0,08	-0,55	-0,03
flws	-0,09	1	-0,02	0	0	-0,02	-0,01	0,01	0,03	0,01	-0,03	0,02	-0,01	0,01
flwg	-0,07	-0,02	1	0,19	0,16	0,01	-0,02	-0,11	0,01	-0,06	-0,12	-0,13	-0,03	0,01
tp	-0,09	0	0,19	1	0,16	-0,14	-0,01	0,03	0,27	0,16	0,01	0,08	0,02	-0,03
ppw	-0,15	0	0,16	0,16	1	-0,02	0,01	0	0,04	0,02	-0,07	-0,06	0,01	0,02
lg	0,09	-0,02	0,01	-0,14	-0,02	1	-0,02	-0,2	-0,68	-0,45	-0,11	-0,28	-0,05	-0,01
nc	-0,01	-0,01	-0,02	-0,01	0,01	-0,02	1	-0,21	-0,04	-0,03	0,07	0,06	-0,02	0
ncm	-0,04	0,01	-0,11	0,03	0	-0,2	-0,21	1	0,28	0,33	0,11	0,22	0,07	-0,01
pcp	-0,18	0,03	0,01	0,27	0,04	-0,68	-0,04	0,28	1	0,79	0,21	0,57	0,14	0,01
uhp	-0,11	0,01	-0,06	0,16	0,02	-0,45	-0,03	0,33	0,79	1	0,2	0,54	0,1	0,03
pc	0,33	-0,03	-0,12	0,01	-0,07	-0,11	0,07	0,11	0,21	0,2	1	0,23	-0,29	-0,08
ppp	-0,08	0,02	-0,13	0,08	-0,06	-0,28	0,06	0,22	0,57	0,54	0,23	1	0,14	-0,01
ps	-0,55	-0,01	-0,03	0,02	0,01	-0,05	-0,02	0,07	0,14	0,1	-0,29	0,14	1	0,17
psi	-0,03	0,01	0,01	-0,03	0,02	-0,01	0	-0,01	0,01	0,03	-0,08	-0,01	0,17	1

where:

m is the number of dimensions for the input
o is the number of dimensions for the output

The implementation was carried out with 100 neurons in the hidden layer and a limit of 1000 iterations the results can be seen in the table IV

5) AdaBoost: is a reinforcement algorithm and presents you with a model in which there are "weak students" (such as small decision trees) [20], their predictions are added to predict the final prediction. this algorithm can be used for both classification and regression problems. The results shown in table IV are quite similar to those obtained by the logistic regression

6) Gaussian Naive Bayes: It is a supervised learning algorithm method based on the Bayes theorem, specifically the Gaussian Naive Bayes algorithm was used that uses the equation 6 which assumes that the probability of the characteristics is Gaussian.

$$P\left(x_i \mid y\right) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\sigma_y^2}\right) \tag{6}$$

where:

The parameters σ_y and μ_y are estimated using maximum likelihood.

7) Quadratic Discriminant Analysis: It is a classic classifier, with simple implementation since they do not have hyperparameters to adjust. depending on the problem they can be a good solution. In the implementation, this algorithm was the one with the worst results due to the non-linearity of some variables (nc, ncm, ps, psi).

8) Gaussian process classification: This unsupervised learning algorithm bases its predictions on Gaussian probabilistic prediction. These algorithms are sensitive to large spaces and lose efficiency, in some implementations the results are not the best [21]. In the implementation this algorithm did not obtain the best, in the prediction of false profiles is where it was most evident. Results shown in table IV.

9) Support Vector Machine: It is an effective supervised learning method used in large spaces, it uses a subset of training points called support vectors. The C-Support Vector Classification (SVC) implementation was used.

$$\sum_{i \in SV} y_i \alpha_i K\left(x_i, x\right) + b \tag{7}$$

The equation 7 the result of the optimization of the classification problem in a hyperplane, this is called a decision function where it is only required to add the support vectors since the coefficients of α_i are zero for the other samples [22].

This implementation is not more convenient in detecting false profiles see table IV

10) Neural Network: Neural network algorithms applied to binary classification problems can be a good implementation as discussed in [23]. Where these algorithms have advantages and disadvantages in implementation, the algorithms that have the best results are the complementary neural networks CMTNN. However, on this occasion, a 4-layer neural network was implemented, see figure 4.

IV. RESULTS

To test the dataset, several machine learning algorithms were used. Mostly classification algorithms showing metrics for accuracy, precision true, precision false. evaluated taken 20% of the data from the dataset as test data. The results shown in table IV, additional the representation of the ROC curve can be visualized in the figure 3, where the fraction of the rate of true positives is represented against the rate of false positives as the discrimination threshold varies.

The algorithm with the best results is Random Forest because it obtained the best accuracy as well as the best true and false prediction precision. It is not surprising since by the selection of characteristics this algorithm is capable of estimating important variables, computing proximity of pairs, locating outliers. The only concern with this algorithm is

TABLE IV
CLASSIFICATION RESULTS (%)

#	Algorithm	Accuracy	A true	A false
1	Decision Tree (DT)	0.92	0.95	0.76
2	Logistic Regression	0.92	0.92	0.88
3	Random Forest	0.96	0.97	0.94
4	Multi-layer Perceptron	0.94	0.96	0.83
5	AdaBoost	0.94	0.95	0.85
6	Gaussian Naive Bayes	0.86	0.86	0.60
7	Quadratic Discriminant Analysis	0.15	0.15	0.00
8	Gaussian process classification (GPC)	0.88	0.89	0.67
9	C-Support Vector	0.88	0.89	0.67

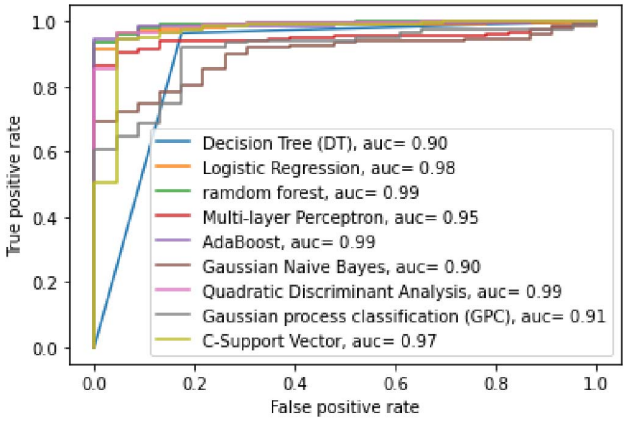

Fig. 3. Comparison of ROC curves for models

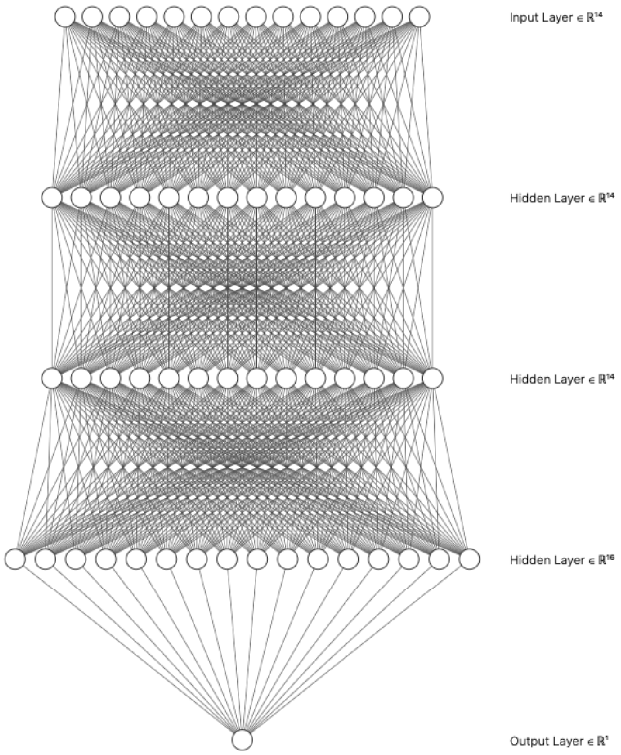

Fig. 4. Architecture neural network

that there are correlated characteristics, small groups are less favored than large groups [24].

Additionally, a test was carried out with a neural network algorithm to verify its behavior with this dataset, specifically a 4-layer neural network was used, its architecture is represented in the figure 4.

This approach generates good results after 40 epochs, executed at 1000 epochs it gives an accuracy of 0.9677 see figure 6, very close to the implementation of algorithms such as Random Forest, Multi-layer Perceptron or AdaBoost. These implementations with these algorithms are also valid approaches that yield good results [23].

With these results and adjusting the parameters, a prediction close to classification algorithms such as Multi-layer Perceptron or AdaBoost can be achieved.

V. CONCLUSIONS

The results of the tests of the classification algorithms such as the implementation of neural networks yielded very good results in the prediction of true and false profiles, giving the best results with Random forest with a 96 percent accuracy followed by Multi-layer Perceptron and the implementation of neural network. The definition of the characteristics of the profile and the analysis of images carried out in the publications are reflected in these results, making this modern research regarding implementations based on metadata. This can be a basis for developing machine learning applications

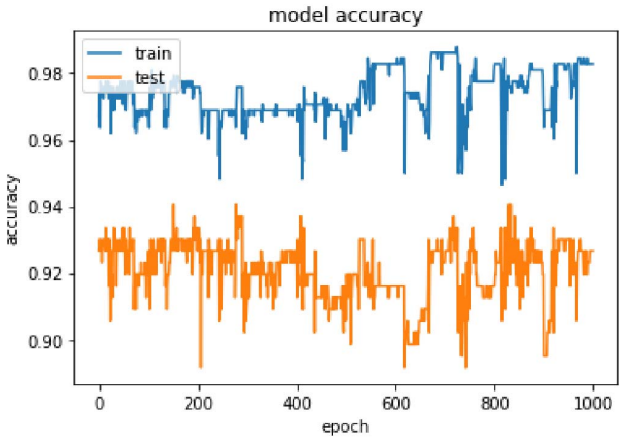

Fig. 5. Epochs – Acurracy neural network result

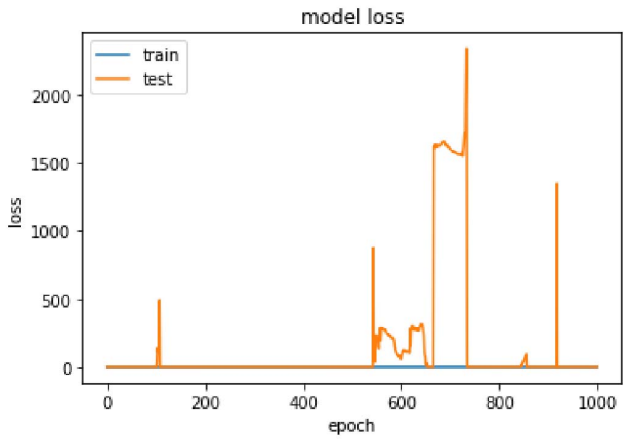

Fig. 6. Epochs – Loss neural network result

that, given a profile, analyze it and predict whether it becomes reliable, helping existing impersonation problems as well as potentially dangerous profiles.

With the tools generated for data extraction, the size of the dataset can be increased to make it more diversified and applicable to more areas. In addition, more features that focus on the content of the publications can be added as comments and from there to an NPL analysis. A more detailed adjustment in the hyperparameters of the algorithms can improve the results obtained.

REFERENCES

[1] M. A. Adebowale, K. T. Lwin, and M. A. Hossain, "Intelligent phishing detection scheme using deep learning algorithms," *Journal of Enterprise Information Management*, vol. ahead-of-print, no. ahead-of-print, Jun. 2020. [Online]. Available: https://doi.org/10.1108/jeim-01-2020-0036

[2] P. Hao and X. Wang, "Integrating PHY security into NDN-IoT networks by exploiting MEC: Authentication efficiency, robustness, and accuracy enhancement," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 5, no. 4, pp. 792–806, Dec. 2019. [Online]. Available: https://doi.org/10.1109/tsipn.2019.2932678

[3] M. Fire, R. Goldschmidt, and Y. Elovici, "Online social networks: Threats and solutions," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 2019–2036, 2014. [Online]. Available: https://doi.org/10.1109/comst.2014.2321628

[4] M. Al-Qurishi, S. M. M. Rahman, M. S. Hossain, A. Almogren, M. Alrubaian, A. Alamri, M. Al-Rakhami, and B. Gupta, "An efficient key agreement protocol for sybil-precaution in online social networks," *Future Generation Computer Systems*, vol. 84, pp. 139–148, Jul. 2018. [Online]. Available: https://doi.org/10.1016/j.future.2017.07.055

[5] M. Foody, M. Samara, and P. Carlbring, "A review of cyberbullying and suggestions for online psychological therapy," *Internet Interventions*, vol. 2, no. 3, pp. 235–242, Sep. 2015. [Online]. Available: https://doi.org/10.1016/j.invent.2015.05.002

[6] statista, "Number of monthly active instagram users," 2020, last accessed on 2020-11-14. [Online]. Available: https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/

[7] A. Sapountzi and K. E. Psannis, "Social networking data analysis tools & challenges," *Future Generation Computer Systems*, vol. 86, pp. 893–913, Sep. 2018. [Online]. Available: https://doi.org/10.1016/j.future.2016.10.019

[8] Z. Zhang and B. B. Gupta, "Social media security and trustworthiness: Overview and new direction," *Future Generation Computer Systems*, vol. 86, pp. 914–925, Sep. 2018. [Online]. Available: https://doi.org/10.1016/j.future.2016.10.007

[9] K. S. Adewole, T. Han, W. Wu, H. Song, and A. K. Sangaiah, "Twitter spam account detection based on clustering and classification methods," *The Journal of Supercomputing*, vol. 76, no. 7, pp. 4802–4837, Oct. 2018. [Online]. Available: https://doi.org/10.1007/s11227-018-2641-x

[10] H. Arshad, E. Omlara, I. O. Abiodun, and A. Aminu, "A semi-automated forensic investigation model for online social networks," *Computers & Security*, vol. 97, p. 101946, Oct. 2020. [Online]. Available: https://doi.org/10.1016/j.cose.2020.101946

[11] R. R. Rout, G. Lingam, and D. V. L. N. Somayajulu, "Detection of malicious social bots using learning automata with URL features in twitter network," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 4, pp. 1004–1018, Aug. 2020. [Online]. Available: https://doi.org/10.1109/tcss.2020.2992223

[12] J. Rodríguez-Ruiz, J. I. Mata-Sánchez, R. Monroy, O. Loyola-González, and A. López-Cuevas, "A one-class classification approach for bot detection on twitter," *Computers & Security*, vol. 91, p. 101715, Apr. 2020. [Online]. Available: https://doi.org/10.1016/j.cose.2020.101715

[13] M. Latah, "Detection of malicious social bots: A survey and a refined taxonomy," *Expert Systems with Applications*, vol. 151, p. 113383, Aug. 2020. [Online]. Available: https://doi.org/10.1016/j.eswa.2020.113383

[14] P. Wanda and H. J. Jie, "DeepProfile: Finding fake profile in online social network using dynamic CNN," *Journal of Information Security and Applications*, vol. 52, p. 102465, Jun. 2020. [Online]. Available: https://doi.org/10.1016/j.jisa.2020.102465

[15] Y. Boshmaf, D. Logothetis, G. Siganos, J. Lería, J. Lorenzo, M. Ripeanu, K. Beznosov, and H. Halawa, "Íntegro: Leveraging victim prediction for robust fake account detection in large scale OSNs," *Computers & Security*, vol. 61, pp. 142–168, Aug. 2016. [Online]. Available: https://doi.org/10.1016/j.cose.2016.05.005

[16] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*, 5th ed., ser. Wiley series in probability and statistics. Hoboken, NJ: Wiley, 2012, no. 821.

[17] J. Quinlan, "Simplifying decision trees," *International Journal of Man-Machine Studies*, vol. 27, no. 3, pp. 221 – 234, 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020737387800536

[18] D. Westreich, J. Lessler, and M. J. Funk, "Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression," *Journal of Clinical Epidemiology*, vol. 63, no. 8, pp. 826 – 833, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0895435610001022

[19] L. Breiman, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://doi.org/10.1023/a:1010933404324

[20] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S002200009791504X

[21] C. Williams and D. Barber, "Bayesian classification with gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1342–1351, 1998. [Online]. Available: https://doi.org/10.1109/34.735807

[22] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004. [Online]. Available: https://doi.org/10.1023/b:stco.0000035301.49549.88

[23] P. Jeatrakul and K. Wong, "Comparing the performance of different neural networks for binary classification problems," in *2009 Eighth International Symposium on Natural Language Processing*. IEEE, Oct. 2009. [Online]. Available: https://doi.org/10.1109/snlp.2009.5340935

[24] L. Tolo̧si and T. Lengauer, "Classification with correlated features: unreliability of feature ranking and solutions," *Bioinformatics*, vol. 27, no. 14, pp. 1986–1994, May 2011. [Online]. Available: https://doi.org/10.1093/bioinformatics/btr300