Fake User Identification on Social Network

K Aditya Srivatsav [1], H Abhignya [2], J Manasa [3,] V Devashekhar [4]
UG Scholar, Guru Nanak Institutions Technical Campus, Hyderabad[1,2,3] Associate Professor, Guru Nanak Institutions Technical Campus, Hyderabad [4]

ABSTRACT:
This paper proposes a detection method for fake and clone profiles on Twitter. Clone profiles are detected using two methods: the C4.5 decision tree algorithm and Similarity Measures (evaluating attributes and network relationships). A comparison of these methods shows their effectiveness in detecting clone profiles. Online Social Networks (OSNs) face growing security and privacy concerns, especially with fake and clone profiles—fake profiles are created for malicious activities, while clone profiles replicate legitimate user information to harm the original owner's identity and launch threats like phishing or spamming.

1 INTRODUCTION

Overview Although Facebook, Instagram, Twitter, and other online social networks (OSNs) are popular for networking, they also present significant security vulnerabilities. Users frequently divulge private information, including images, contact information, and affiliations, which can have serious repercussions if misused by hackers. A serious risk is profile cloning, which is the process of making phony accounts with user information that has been stolen. This allows attackers to pose as users, disseminate false information, and carry out harmful actions. It can happen on the same platform (same-site cloning) or on separate platforms (cross-site cloning). Because social network registration has become easier, more people are creating phony profiles, which has led to problems like phishing, cyberbullying, and spam. Advanced detection techniques are desperately needed to counter these threats in order to protect user identity and improve

Social networks offer diverse benefits for organizations by enhancing learning, collaboration, and communication. They support informal learning by connecting groups of learners and fostering social connections, while also aiding all organizational members in building communities of practice. Social networks enable engagement by providing business intelligence and feedback, though ethical concerns may arise. Their ease of use simplifies access to tools and applications, as seen with platforms like Facebook. Additionally, a familiar interface spanning work and social boundaries minimizes the need for training, though it may blur lines between personal and professional activities.

1.1 OBJECTIVE

This research proposes a detection algorithm that can identify Twitter clones and fake profiles. A collection of rules that can successfully distinguish between real and false profiles is used to detect fake profiles. Two techniques are employed to detect profile cloning. One employs the C4.5 decision tree algorithm, while the other uses similarity measures. Similarity of attributes and similarity of network linkages are the two categories of similarities taken into account in similarity measures. C4.5 uses a decision tree that takes information gain into account to identify clones. The effectiveness of these two approaches in identifying clone profiles is evaluated through comparison.

1.2 SCOPE OF THE WORK:

The societal hazard posed by fake and clone profiles has grown significantly. Hackers can simply establish fake or clone profiles using information that is easily accessible on social networks, such as phone number, email address, school or college name, company name, location, etc. They then attempt to launch several types of attacks, such as cyberbullying, spam, and phishing. They even attempt to discredit the organisation or its rightful owner. In order to improve the security of users' social lives, a detection technique that can identify both phoney and clone profiles has been developed.

1.3 PROBLEM STATEMENT

when user credentials are obtained from one network and used to build a clone profile in the same network. By using user data from one network to generate a duplicate profile on another network where the user does not have an account, the attacker is engaging in cross-site profile cloning. Fake profiles are growing at an alarming rate as social networks have made the registration procedure extremely easy in an effort to draw in more members. To connect with a target and carry out malicious actions, an attacker fabricates a profile. Additionally, to disseminate spam and bogus news.

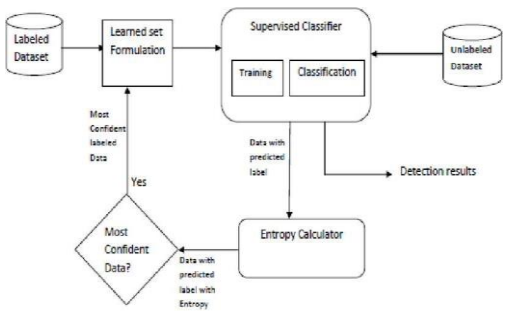
1.4 EXISTING SYSTEM:

Numerous techniques have been put forth by researchers to identify phoney accounts and duplicated profiles on social media sites. Two methods for identifying cloned profiles were presented by Brodka, Mateusz Sobas, and Henric Johnson. One method looks at network relationships, while the other depends on how similar the attribute values of the original and cloned profiles are. Profiles with the same name are found and their similarity is computed by using the victim's name as a primary key. The user manually confirms the validity of the profile once it is marked as a possible clone if the resemblance over a certain threshold. Similar to this, Crecci et al. examined important characteristics and guidelines for identifying phoney Twitter accounts. They trained machine learning classifiers using these features, producing a Class A classifier that can successfully discriminate between authentic and fraudulent accounts.

1.4.1 Existing System Disadvantages:

- The threat posed by fake and clone profiles to society has grown significantly. Hackers can simply establish fake or clone profiles using information that is easily accessible on social networks, such as phone number, email address, school or college name, company name, location, etc.
- They then attempt to launch several types of attacks, such as cyberbullying, spam, and phishing. They even attempt to discredit the organisation or its rightful owner.

2 SYSTEM ARCHITECTURE:



2.1 EXPLANATION:

The document's system architecture is intended to identify phoney and clone profiles on social media platforms such as Twitter. Fake Profile Detection and Clone Profile Detection are its two primary modules. When profiles surpass a predetermined level, the Fake Profile Detection module classifies them as fake based on a set of established rules, including the absence of profile names or photographs, a lack of description, deactivated geolocation, and odd tweeting habits. The Clone Profile Detection module uses two methods: Network Similarity, which examines links and linkages within the network to identify cloning trends, and Attribute Similarity, which compares profile attributes like name and email.To ensure precise detection and improved security, profiles that over a predetermined threshold are marked as clones based on a similarity index.

2.2 PROPOSED SYSTEM

The suggested method for identifying phoney Twitter identities employs a rule-based methodology to pinpoint questionable accounts. Important clues include the absence of an account description, a profile name or image, and a geo-enabled field that is disabled and frequently used to conceal location data. Red flags could include irregular activity patterns, such as an abnormally high volume of tweets or none at all. A counter is increased for each rule found, and profiles that surpass a predetermined threshold are flagged as fraudulent. A safer online social environment is promoted by this method's systematic evaluation of profiles, which improves the precision and dependability of fake profile detection.

2.2.1 PROPOSED SYSTEM ADVANTAGES:

- The modules worked fine and was able to detect clones with good accuracy.
- Good Results

3 DESCRIPTION

3.1 GENERAL:

In the initial step, they selected and weighted a few useful features for the detection process that they had gathered from other studies. To find the minimal set of characteristics that yield correct findings, numerous experiments are carried out. Only seven characteristics—out of 22—were chosen because they are capable of identifying fraudulent accounts, and these characteristics have been used in classification methods. The most accurate categorization method is chosen after a comparison of the methods is conducted based on the results.

3.2 METHODOLOGIES

3.2.1 MODULES NAME:

- Data Collection
- Dataset
- Data Preparation
- Model Selection
- Analyze and Prediction
- Accuracy on test set

Data Collection

Data collection involves systematically collecting, evaluating, and interpreting information using established methods to ensure accuracy and validity. Researchers rely on this data to verify their hypotheses. As a fundamental and initial phase in any research project, data collection is indispensable regardless of the field. The techniques for collecting data differ among various disciplines based on the specific information required.

Dataset

The dataset used in the system comprises 1,338 individual records, with each entry represented by nine columns of information. These columns include: **ID**, a unique identifier for each record; **UserID**, representing the Twitter account ID; **No Of Abuse Report**, indicating the number of abuse reports associated with the account; **No Of Rejected Friend Requests**, the count of declined friend requests; **No Of Friends**, the total number of friends; **No Of Followers**, the total number of followers; **No Of Likes To Unknown Account**, capturing the number of likes given to unrecognized accounts; **No Of Comments Per Day**, indicating the daily average number of comments; and Fake Or Not Category

Data Preparation

We'll change the data. by eliminating certain columns and missing data. We will start by compiling a list of the column names that we wish to preserve.
After that, we eliminate or drop every column save for the ones we choose to keep. Lastly, we eliminate from the data set the rows that have missingvalues.

Model Selection

The dataset is divided into two sections: 80% for training and 20% for testing in order to build a machine learning model. The `train_test_split` tool from scikit-learn is used to accomplish this, splitting the dataset into feature and label columns. The split ratio is specified by the `test_size` parameter, and repeatability is guaranteed by seeding the random number generator with the
`random_state` option. Four datasets are produced by the function: `train_x`, `train_y`, `test_x`, and `test_y`.

A Random Forest Classifier is trained by utilizing its ability to fit multiple decision trees to the data. The training process involves feeding train_x and train_y into the fit function. Once the model is trained, it is evaluated using the predict method with test_x. As a powerful supervised learning method, Random Forest operates in two phases: first, it builds a collection of decision tree while training, and then it generates predictions by aggregating the results from these trees, making it an effective tool for classification tasks.

Analyze and Prediction

In the dataset used for analysis, we selected seven key features: **UserID**, which represents the Twitter account ID; **No Of Abuse Report**, indicating the number of abuse reports associated with the account; **No Of Rejected Friend Requests**, showing how many friend requests were rejected; **No Of Friends**, representing the number of friends the account has; **No Of Followers**, reflecting the number of followers the account has; **No Of Likes To Unknown Account**, representing the number of likes given to unknown accounts; and **No Of Comments Per Day**, which tracks the average number of comments made by the account daily.

Accuracy on test set

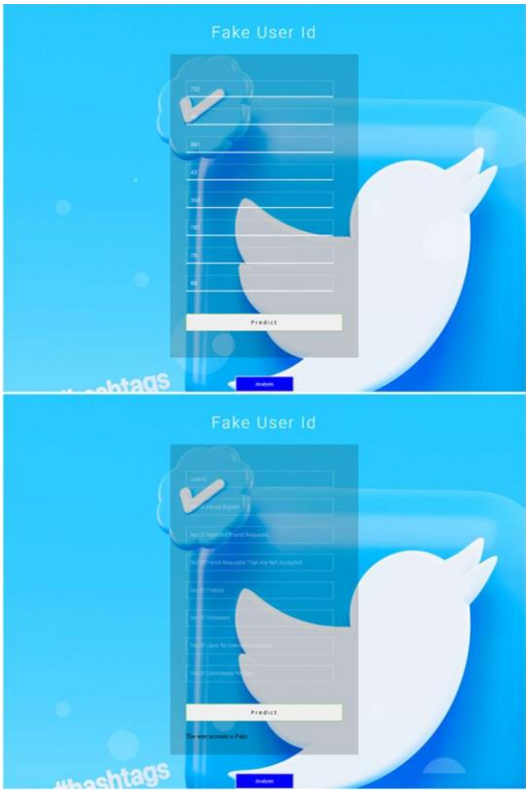
We got a accuracy of 95.1% on test set.

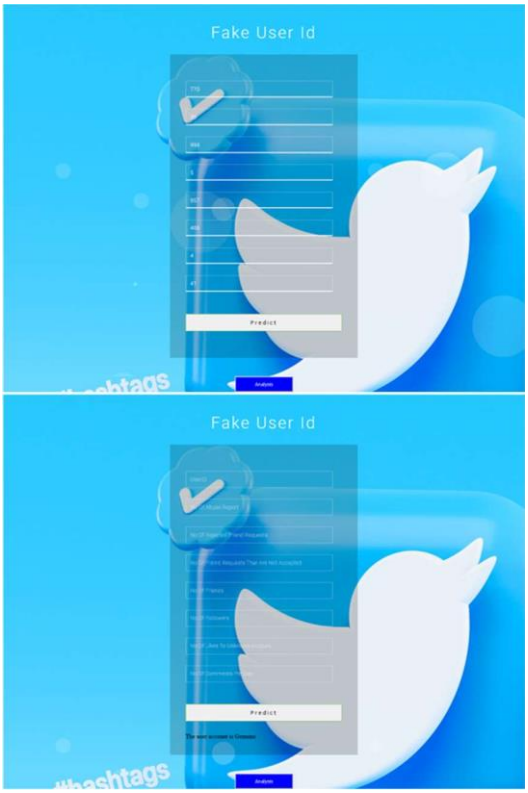
3.3 TECHNIQUE USED OR ALGORITHM USED

DECISION TREE:

Techniques and algorithms for identifying phoney and clone profiles on social networks, particularly Twitter, are covered in the document. The suggested approach combines two methods for clone detection: *Similarity Measures* and the
C4.5 decision tree algorithm. It also uses rule- based detection for false profiles. The Similarity Measures technique computes similarity indices by analysing properties and network interactions. Clone flags are applied to profiles that surpass a specific similarity level. By building a tree using the information gained from profile traits, the C4.5 decision tree method categorises clone profiles. The project also uses features like friend request rejections and abuse reports to train a Random Forest Classifier for prediction tasks. The methods guarantee a methodical detection procedure that accurately identifies rogue accounts.

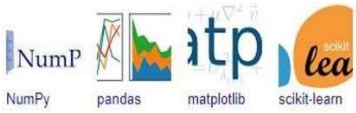
4 RESULT:





Libraries used in python:

- numpy - mainly useful for its N- dimensional array objects.
- pandas - Python data analysis library, including structures such as data frames.
- matplotlib - 2D plotting library producing publication quality figures.
- scikit-learn - the machine learning algorithms used for data analysis and data mining tasks.



5 FUTURE ENHANCEMENT

Only the profile attributes for detecting fakes and clones have been taken into consideration in this work. This study can be expanded in the future by using some NLP approaches to evaluate tweets as well.

6 CONCLUSION

In online social networks, fake and clone profiles have grown to be a major issue. In daily life, we learn about some of the hazards posed by these types. Therefore, a detection technique that can identify phoney and clone Twitter profiles has been proposed. A collection of rules that, when applied, can distinguish between real and fraudulent profiles was employed for fake detection. Similarity Measures and the C4.5 algorithm were used to detect clones, and the results were compared to assess performance. The majority of the clones that were given into the system could be detected using Similarity Measures, which performed better than C4.5.

7 REFERENCES

[1] Sowmya P and Madhumita Chatterjee ," Detection of Fake and Cloned Profiles in Online Social Networks", Proceedings 2019: Conference on Technologies for Future Cities (CTFC)

[2] Georgios Kontaxis, Iasonas Polakis, Sotiris Ioannidis and Evangelos P. Markatos, "Detecting Social Network Profile Cloning", 2013

[3] Piotr Bródka, Mateusz Sobas and Henric Johnson, "Profile Cloning Detection in Social Networks", 2014 European Network Intelligence Conference

[4] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angello Spognardi, Maurizio Tesconi, "Fame for sale: Efficient detection of fake Twitter followers", 2015 Elsevier's journal Decision Support Systems, Volume 80

[5] Ahmed El Azab, Amira M Idrees, Mahmoud A Mahmoud, Hesham Hefny, "Fake Account Detection in

Twitter Based on Minimum Weighted Feature set", World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering Vol:10, 2016

[6] M.A.Devmane and N.K.Rana, "Detection and Prevention of Profile Cloning in Online Social Networks", 2014 IEEE International Conference on Recent Advances and Innovations in Engineering

[7] Kiruthiga. S, Kola Sujatha. P and Kannan. A, "Detecting Cloning Attack in Social Networks Using Classification and Clustering Techniques" 2014 International Conference on Recent Trends in Information Technology

[8] Buket Erşahin, Ozlem Aktaş, Deniz Kilinç, Ceyhun Akyol, "Twitter fake account detection", 2017 International Conference on Computer Science and Engineering (UBMK)

[9] Arpitha D, Shrilakshmi Prasad, Prakruthi S, Raghuram A.S, "Python based Machine Learning for Profile Matching", International Research Journal of Engineering and Technology (IRJET), 2018

[10] Olga Peled, Michael Fire, Lior Rokach, Yuval Elovici, "Entity Matching in Online Social Networks", 2013 International Conference on Social Computing

[11] Aditi Gupta and Rishabh Kaushal, "Towards Detecting Fake User Accounts in Facebook", 2017 ISEA Asia Security and Privacy (ISEASP)

[12] Michael Fire, Roy Goldschmidt, Yuval Elovici, "Online Social Networks: Threats and Solutions", JOURNAL OF LATEX CLASS FILES, VOL. 11,
NO. 4, DECEMBER 2012, IEEE Communications Surveys & Tutorials

[13] Ashraf Khalil, Hassan Hajjdiab and Nabeel Al- Qirim, "Detecting Fake Followers in Twitter: A Machine Learning Approach" 2017 International Journal of Machine Learning and Computing.