JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

IDENTIFYING FAKE PROFILES USING ANN

Mr. Ch. Vijaya Kumar[1], E. Rupa Reddy [2], Nigam Archana [3], S. Ruthik Reddy[4]

ABSTRACT

In sophisticated persistent threats like cyber espionage, including the theft of secrets, fake identities or profiles play a significant role. These are additionally engaged in harmful and risky activities. Currently, social networks are very important for users to carry out daily tasks. A variety of scammers are drawn to social media networks because of their excessive use. These con artists adopt numerous false identities in order to commit various frauds. Using Deep Learning-ANN, we are determining whether the Facebook details are genuine or not in this project. We have taken the Facebook dataset from Github in order to carry out these processes.

False identities or profiles play a crucial part in sophisticated persistent threats like cyber espionage, especially the theft of secrets. These also take part in dangerous and risky activities. Social networks are now crucial for users to complete daily chores. Due to their widespread use, social media networks attract a variety of scams. These con artists use a variety of phoney identities to commit different types of fraud. In this study, we are determining the veracity of the Facebook details using Deep Learning-ANN. To do these operations, we used the Facebook dataset that was available on Github.

Keywords:

Identification of fake accounts (profiles), deep learning artificial neural networks, and machine learning classifications using SVM and ANN.

INTRODUCTION

Every person's life has undergone a transformation thanks to social media. It established its own standard. Social media is a vital part of every person's daily existence. Social media is a platform that connects people all over the world. It links everyone from all corners of the globe. By connecting to the internet, we can contact with everyone using social media. Social media operates online. It produced the idea that the world would end if social media disappeared.

Internally, social media set itself different in several domains. Many businesses use social media to market their goods or services. Numerous phoney profiles are developed because so many things rely on social media platforms like Facebook, Instagram, WhatsApp, Twitter, etc. Many of the millions of Facebook, WhatsApp, and other social media users are fraudulent profiles. This false profile raises a number of security concerns. Someone makes fake profiles to seduce users and defraud them. Through the creation of false profiles, they steal the personal data.

Fake profiles can lead to all kinds of terrible things. User IDs are used to create multiple accounts. They damage the victim's reputation by disseminating false information. They even ask the friends of the victim for financial assistance. Even victim suicides result from this. Therefore, it's crucial to spot the bogus profiles. We can protect the victims by figuring out the phoney profile.

We developed a method to spot the false profiles that makes use of artificial neural networks. Using the Python computer language, we have developed an algorithm to help us spot false profiles. This algorithm might be able

to tell whether a friend request is genuine or not. By training the data set, it is possible for many social media businesses to distinguish between legitimate accounts and false identities.

EXISTING SYSTEM

•False profiles are made by malicious individuals in an effort to steal login credentials from unwary users. Many people with public accounts will get friend requests from a bogus profile. These fake accounts entice unwary viewers with images of persons who are deemed appealing. The owner of the fake profile will spam friend invitations to everyone this person is friends with once the user accepts the request. • Links in the phoney profile's text frequently go to external websites where harm is done. When an uninformed person clicks the harmful link, their PC will be harmed. It may cost as little as contracting a virus to as much as having a rootkit installed that turns the machine into a zombie.

NLP PRE-PROCESSING

Text pre-processing is a crucial component of any NLP method, and the importance of the NLP pre-processing are as follows: 1. To reduce the indexing (or knowledge) records dimension of the textual content records, where i. stop words account for 20-30% of total phrase counts in a particular text, and ii. stemming may only reduce indexing size by 40–50%. 2. To increase the IR method's efficacy and efficiency. Stop words aren't useful for searching or mining textual content, therefore they could merely confuse the retrieval system ii. Stemming is a technique used to match similar words in a text record.

TOKENIZATION:

Tokenization is the process of separating a body of text into tokens, which can be words, phrases, symbols, or other meaningful elements.The examination of the phrases in a sentence is the goal of the tokenization process. The list of tokens serves as the input for subsequent processing, such as text mining or parsing. Tokenization is useful in both linguistics, where it is a method of text segmentation, and computer science, where it is a method of lexical analysis. The beginning of textual knowledge is only a block of characters. The words from the data set are necessary for every method of knowledge retrieval. Because of this, tokenization of records is necessary for a parser. Given that the material has already been recorded in computer-readable codes, this may seem straightforward. However, certain issues are still present, such as the removal of punctuation. Additionally, special characters like brackets and hyphens must be processed.

STOP WORD REMOVAL:

Stop words are used far more frequently than old-fashioned words like "and," "are," "this," etc. They don't appear to be helpful in categorising records. Thus, they must be taken out. However, there are issues with the construction of such stop phrase records and there is inconsistency between textual sources. Additionally, this procedure lowers text knowledge and enhances the effectiveness of the strategy. These words are included in every text content report even though they are not necessary for text mining software.

STEMMING AND LEMMATIZATION:

Scaling down inflectional types and mostly derivationally linked variety of a phrase to a fashioned base kind is the goal of both stemming and lemmatization. In order to achieve this aim more frequently than not properly, stemming typically refers to a primitive heuristic process that removes the endings of words. This procedure frequently entails removing derivational affixes. Lemmatization is a term used to describe the skillful completion of tasks using a vocabulary and morphological analysis of phrases. It typically aims to remove just inflectional ends and restore the lemma, or dictionary type, of a word.

SUPPORT VECTOR MACHINE (SVM):

Finding the exceptional hyperplane that divides all information aspects of one type from those of the other categorization is how an SVM classifies data. The hyperplane with the longest line connecting the two classes is the optimal hyperplane for an SVM technique. Finding the exceptional hyperplane that divides all knowledge aspects of one category from those of the other is how an SVM classifies data. The informational components that are closest to the separating hyperplane are the help vectors.

NAÏVE BAYES :

The Naive Bayes algorithm is an algorithm that determines the likelihood that an item with specific qualities belongs to a particular group or category. It's a probabilistic classifier, to put it briefly. Because it holds that the occurrence of a particular feature is independent of the predominance of other factors, the Naive Bayes algorithm is referred to as "naive". As an example, let's say we want to identify fraudulent profiles based on their location, language, and posts' time, date, or publication. All of these characteristics, in my opinion, increase the likelihood that the false profile will exist, even if they all depend on one another or on the existence of other elements.

PROPOSED SYSTEM :

In this project, artificial neural networks are used to determine if the social network account information provided is from real or bogus users. When we provide new test data, the ANN train model is applied to the new test data to determine if the provided new account information are from real or false users. The ANN algorithm is trained using all previous users' fake and authentic account datasets. Utilising ANN to Spot Fake Profiles Online social networks like Facebook and Twitter store user information, and bad individuals may hack social network databases to steal or otherwise compromise user information. To safeguard user data, we employ the ANN algorithm.

We are leveraging the information from social networks below to train the ANN algorithm. Location, Location_IP, Friend Count, Status, Account_Age, Gender, User Age, Link Description, Status_Count, and User Age The main goal of all fake users is to send friend requests to regular users in order to hack into their computers or steal their data. They never have a large number of posts or friends they follow, and their account ages also have a small number of years. Facebook will determine if a user profile is false or real by studying this feature.

We collected the Facebook profile data from the Facebook website and used it to train an ANN model. Some values from the profile dataset are shown below.

Location, Location_IP, Friend Count, Status, Account_Age, Gender, User Age, Link Description, Status_Count, and User Age

10, 1, 22, 0, 1073, 237, 0, 0, 0

10, 0, 33, 0, 127, 152, 0, 0, 0

10, 1, 46, 0, 1601, 405, 0, 0, 0

10, 0, 25, 0, 704, 380, 0, 0, 0

7, 1, 34, 1, 64, 721, 1, 1,

1 7, 1, 30, 1, 69, 587, 1, 1,

1 7, 1, 36, 1, 61, 782, 1, 1,

1 7, 1, 52, 1, 96, 827, 1, 1, 1

All bold names in the aforementioned dataset are the names of its columns, and all integer numbers are the dataset's values. Due to ANN's inability to handle string values, we transform gender values to 0 or 1, depending on whether they are male or female. In the dataset above, the last column indicates whether an account is real or phoney. If the last column has a value of 0, the account is real; otherwise, it is false. Since sending friend requests, rather than making posts, is the primary goal of all fake accounts, they will all have less posts. By examining these qualities, Facebook marks the record with value 1, indicating that it is a fake account.We are utilising the aforementioned dataset, which is kept within the code folder named "dataset," to train an ANN model. After creating the train model, test data and account information are entered, and an ANN determines if the outcome is phoney or real. Here are some test data values: Location_IP, Location, Status_Count, Friend_Count, Account_Age, Gender, User Age, Link_Desc Utilising ANN to Spot Fake Profiles 9 10, 1, 44, 0, 280, 1273, 0, 0 10, 0, 54, 0, 5237, 241, 0, 0 7, 0, 42, 1, 57, 631, 1, 1 7, 1, 56, 1, 66, 623, 1, 1 The STATUS column in the test data above lacks a value, and an ANN will predict the status and tell us if the test data is real or not. The output of the above test data will indicate if it was real or not.

PERFORMANCE EVALUATION:

The specific statistic and the issue at hand must be taken into consideration when interpreting performance evaluation metrics. Higher levels of some indicators, however, typically imply greater performance, whereas lower ones typically suggest inferior performance. A broad guideline for frequently employed performance evaluation measures is provided below:

- **Accuracy**: It evaluates how accurately the model or method's predictions were produced overall. It is determined as the proportion of occurrences that were properly categorised to all instances.

• **Mean Absolute Error**: It speaks about the size of the discrepancy between the observation's predicted value and its actual value. The magnitude of errors for the entire group is determined by averaging the absolute errors for a set of forecasts and observations.

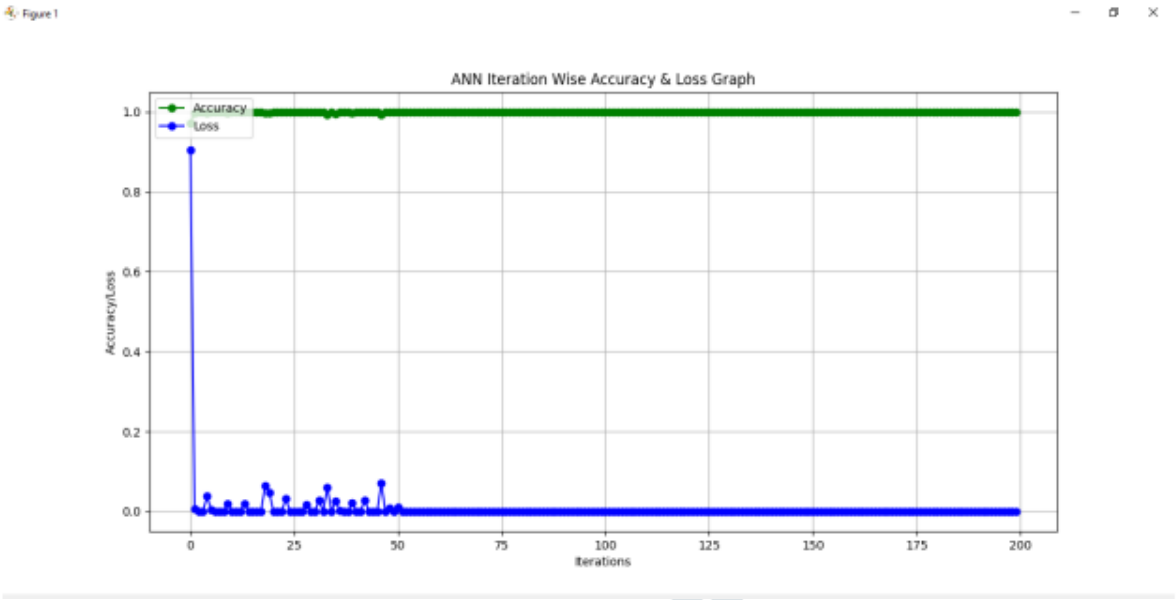
• **Root Mean Squared Error**: It calculates the typical difference between values that a model predicts and actual values. It gives an estimate of the model's accuracy in predicting the desired value.

• **Loss**: It is a fine for making a poor prediction. In other words, loss is a measure of how poorly the model predicted a particular case.

Evaluation performance of ANN:

The following metrics are employed in the performance analysis of the ANN: Accuracy is the ratio of correct predictions to all predictions. Mean Absolute Error (MAE) is determined by adding

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \vartheta(i)\|^2}{N}},$$



divided by the sample size, in terms of absolute errors ((1/n) (i=1 to n) |y_i - _i|) is the actual value. Calculate the residual (difference between the forecast and the truth) for each data point, together with its norm and mean, and then take the square root of that mean to arrive at the root mean squared error (RMSE).

CONCLUSION

The goal of the project is to develop an artificial neural network by continuously collecting newspaper articles. When we access or log into social media, we receive numerous friend requests. A request might be genuine or fake. The identification of false accounts or profiles using artificial neural networks is thoroughly explained in this study.

ACKNOWLEDGEMENT

We appreciate Mr. Ch. Vijay Kumar and Mrs. Soppari Kavitha for their important time and advice as our guide. Additionally, Dr. M. V. VIJAYA SARADHI, Head of the Computer Science and Engineering Department, is much appreciated for his support of Ace Engineering College and unfailing time.

REFERENCES

[1] Social Media - Statistics & Facts, S. Dixon, June 21, 2022.

[2] Political Advertising Spending on Facebook between 2014 and 2018 by Sponsor Category, Anita Balakrishnan and A. Guttmann, September 18, 2018. In Q4, which ended on January 31, 2018, Facebook made an average of $6.8 from each user.

[3] R. Nieva and L. Hautala 50 million Facebook users' data were at danger due to a breach on September 28, 2018

[4] Xiaowei Yang, Tiago Pregueiro, Qiango Cao, Michal Sirivianos, Making it easier to spot false accounts on large-scale social networking sites, Conference: Networked Systems Design and Implementation, 9th USENIX Conference, April 2012.

[5] A Mishra and Akshay J. Sarode published Audit and Analysis of Impostors: "An experimental approach to detect fake profile in online social network" in the 2015 ICCCT Proceedings, pp. 1–8.

[6] Fake profile detection methods in large-scale online social networks: A thorough analysis, D. Ramalingam and V. Chinnaiah, Computers & Electrical Engineering, vol. 65, p. 177, 2018.

[7] List of Top Countries with the Most Cybercrime According to Sumo 3000 in Computer Security, Enigmasoftware.com

[8] R Kaur and S Singh published "A survey of data mining and social network analysis based anomaly detection techniques" in the Egyptian Informatics Journal, Vol. 17, No. 2, pp. 199–216, in December 2015.

[9] the duo of Akshay J. Sarode and Arun Mishra. 2015's IRE Journals article, "Using Facebook Graph API Tool,"