International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 1027-1032 www.ijaem.net ISSN: 2395-5252

An Approach for Detecting a Fake Account using Supervised Learning Algorithm

Bhargavi Vegesna1, J Jithin Ambala2, Jagadeesh ReddyAbbireddy3, Om prudhvi raj Jatla 4

1-4UGStudent, Dept. of Computer science Engineering, GITAM UNIVERITY, Visakhapatnam, Andhra Pradesh, India.

Date of Submission: 12-04-2023 Date of Acceptance: 22-04-2023

ABSTRACT:
At Online social networks (OSNs) are more and more common, which affects people's social life and encourages them to sign up for different social media platforms. Several operations, including promotion, communication, agenda creation, advertisement generation, and news creation, are now carried out primarily on social media platforms. Making new friends, staying in touch with them, and staying up to date on their updates have all been simpler. Social media is used by small businesses to advertise themselves. In order to understand how these online social networks, affect people, researchers have started examining them.
Some malevolent accounts are used to spread false information, commit fraud, and advance political agendas. An important step is the detection of fraudulent accounts. Machine learning-based techniques were employed to identify fictitious accounts that might deceive users. The dataset is pre-processed using several Python libraries, and a comparison model is obtained to find an efficient solution that works with the provided dataset. Several Machine Learning techniques are used to try and identify fake accounts on social media platforms. The methods Support Vector Machines and Random Forest's classification capabilities are utilized to identify duplicate accounts.

I. INTRODUCTION:

On a social networking site, users can build a profile, stay in touch with friends, publish updates, and connect with new people who have similar interests. These Online Social Networks (OSN) leverage Web2.0 technology to link people with one another. The use of social networking sites is growing and evolving quickly. Users can more easily make new friends thanks to online groups that bring individuals together with shared interests. Social networks are now the most pervasive aspect of everyone's lives in the modern world due to their increased impact.

There are many uses for social networking, including business, employment, education, and shopping. It has advantages and disadvantages because it has gained such popularity. The optimal algorithm between random forest and support vector machine that can assess the likelihood that a social media profile is authentic or false will be found in order to avoid these shortcomings and ensure that users are safe while using social media. Datasets of fictitious users and real users will be collected from Kaggle and divided between training and testing portions with 80% and 20%, respectively. Simple imputer will be used to forecast the missing values in the dataset, aiding in model training.The model will be tested after being trained using the support vector machine and random forest methods. Confusion matrix is used to calculate accuracy, or the capacity to determine if an account is false or real.

II. LITERATURE SURVEY:

[1] Vijay Tiwari, Ministry of defence proposed Analysis and Detection of Fake Profile Over Social Network: In this paper on bot detection is done using three methods bot detection based on scrutiny of content, detection based on network graph and combination or hybrid approach Machine learning methods used in this paper was post method, accuracy detection models and supervised learning.
[2] in this paper Detecting Fake Account on social media: In this paper neural networks and support vector machine are used and within its algorithm data pre-processing is done after that in feature reduction principal component analysis, Spearman's Rank-Order Correlation, Relevance and Redundancy Analysis Technique, Markov Blanket Technique, Wrapper Feature Selection

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 1027-1032 www.ijaem.net ISSN: 2395-5252

using SVM are used and SVM-NN classification are done.

[3] in this statement Facebook said that there almost 4.3 percent of its active user accounts are duplicate, and almost 83 million user accounts are fake which is increasing extremely fast.

[4] In this paper discus that social media growing extremely fast in field of entertainment, and business we use social media in every field but this all-social media platform having some issue like trolling, bulling, fraud mostly this done by using fake accounts.

[5]Ali M. Meligy (2017) This paper presents a technique to detect fake accounts on social networking site called fake profile recognizer. This technique is based on two methods i.e regular expression and deterministic finite automata. A regular expression is used to authenticate the profiles and deterministic automata recognize the identities in trusted manner.

III. PROBLEM IDENTIFICATION

3.1 Project Definition:

Thisprojectaimsto determine the best accurate and efficient algorithm between the Random Forest method and Support Vector Machine to determine whether Facebook, Twitter, and Instagram accounts are fake or real.

3.1.1 Systems that are already inuse:

The detection of false accounts on online social networking platforms is a significant scientific challenge. Users of online social networks are constantly confronted with security-related problems, which have an impact on both their social and personal lives. A social network is seeing a daily rise in the number of fraudulent accounts. Fake ratings, promote spam, and distribute fake news through accounts. Identifying fake accounts on social media websites is the goal of our proposed strategy. It is possible to identify fake accounts on online social networks using a number of different ways. Everybody has different goals, benefits, and drawbacks. It has been found through research that not all of the currently used approaches have very high f-measure and recall values.

3.1.2 PROPOSED SYSTEM:

For the purpose of classifying authentic accounts from false ones, the suggested work employs techniques like random forest and support vector machines, or SVM. On social media websites, the feature set that affects the identification of bogus accounts will be employed. This suggested effort is anticipated to produce the higher f-measure value and recall needed for fraudulent account identification on social media networks. The results produced by the machine learning methods of random forest and support vector machine are precise. We can choose the best algorithm to find fake accounts based on accuracy.

IV. SYSTEM ARCHITECTURE.

As shown in figure (1) there is a database that consists of all the Social Media Profiles which we have considered. This module is trained repetitively to obtain maximum accuracy using a classification algorithm. If a new profile is given to the module, it will classify whether the given profile is fake or not. And then provides the appropriate result.

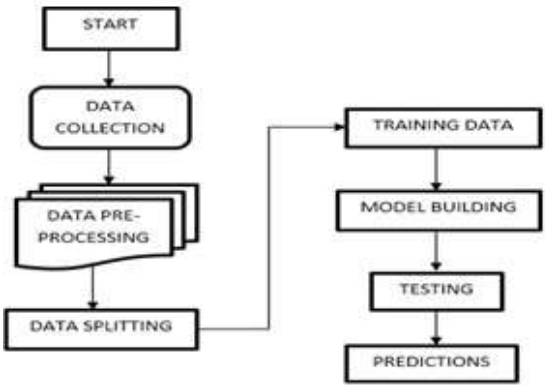

Figure 1: System Architecture

4.1 Proposed flow chart:

Figure2: Data flow diagram of the processes involved.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 1027-1032 www.ijaem.net ISSN: 2395-5252

4.2 UMl Representation:

An illustration of class shows how the physical parts of a system are wired together and organized. It is common practice to create deployment diagrams in order to model implementation details and confirm that all necessary functions of the system are addressed in the development plan.

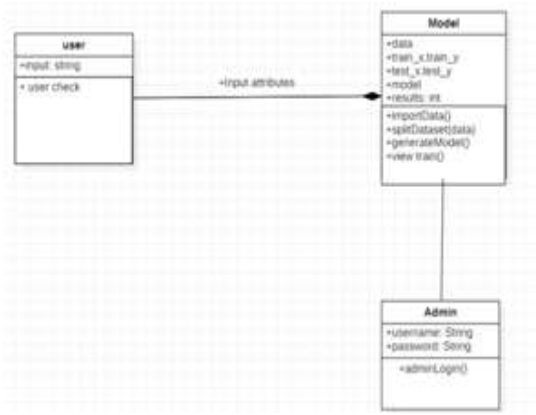

Figure 3: Class diagram **detection of faces (HOG-histogram of gradients)**

- The first step is to convert the image into black and white images, as we don't require color image for face detection.
- Next, for each pixel, we will be getting a gradient from the darker region to the lighter region based on the gradient value in x direction and y direction and tan inverse of x gradient by y-gradient gives us the direction of the gradient.
- After every pixel gets its gradient value, we match the image with a pre- configured image so that we can tell whether there is a face or not
- The above steps are repeated for all the faces in an image so that all the faces are identified
- This step is only for detecting the faces in the image

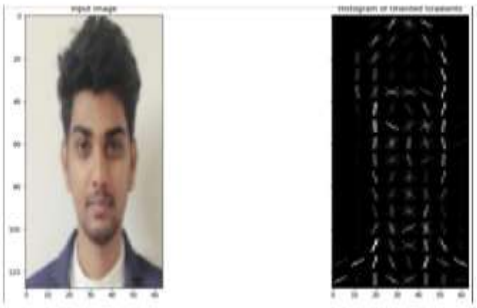

Figure-2:Histogram of oriented gradients

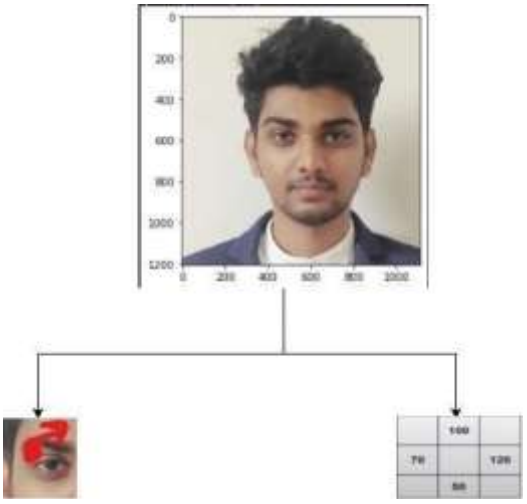

Figure3. How each pixel gradient is obtained
Gradientin x-direction=120-70=50
Gradient in y-direction=100-50=50
Gradient magnitude=square root((50)^2+(50)^2))=70.1Gradientangle=tan inverse(50/50)=45degrees

4.3 Software Environment:

a. Python

Python is a sophisticated scripting language that can be interactively used and is object-oriented. High reading comprehension is intended. In contrast to other languages, it typically uses English keywords and has fewer syntactic structures.

b. Google colab

Google Research created Colab as a product. Because it allows anyone to write and run arbitrary Python code through the internet, Colab is incredibly helpful for machine learning, data analysis, and education. In a technical sense,Colab offers unrestricted access to computer resources, including GPUs, and is a hosted Jupyter notebook service.

4.4 Overview of Technologies:

a. PYTHON - Preferred programming language

Python is one of the programming languages that is easiest for beginners to learn due of its straightforward syntax.
Python's benefits include:
• Python community that is mature and supportive
• Simple to Learn and Use
• Numerous Python Frameworks and Libraries
• Usage in Big Data, Machine Learning, and Cloud Computing
• Versatility, Efficiency, Reliability, flexibility and Speed

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 1027-1032 www.ijaem.net ISSN: 2395-5252

b. Modules in python:

1. PANDAS:

Thanks to pandas, a Python programme that provides rapid, adaptive, and expressive data structures, working with "relational" or "labelled" data can be easy and uncomplicated. It hopes to act as Python's fundamental, high-level building block for carrying out beneficial, practical data analysis in the actual world.

2. NUMPY:

For array manipulation, utilise the NumPy Python module.

It also offers functions for working in the field of linear algebra as well as matrices and the Fourier transform.

Although lists take a while to run, they are Python's counterpart of arrays.

By providing array objects that are up to 50 times faster than typical Python lists, NumPy aims to achieve this. The NumPy array object is called ndarray, and it contains a variety of supporting functions that make using ndarray quite easy.

c. Algorithms:

1. Random Forest:

The Random Forest classifier makes use of a number of decision trees on various subsets of the input data to improve the predicted accuracy of the dataset.

2. Super Vector Machine (SVM):

The goal of the SVM algorithm is to find the optimal decision boundary or line that can classify the n-dimensional space into subspaces, allowing following data points to be swiftly classified. This ideal decision boundary is referred to as a "hyperplane".

V. DATA COLLECTION:

The dataset that was utilized in this to train and test the model was obtained from the open-source website Kaggle. There are 33 attributes in total in the dataset. The profile data is divided into two different files, one for fake profiles and the other for real ones. There are 1338 instances in the first file and 1482 instances in the second file. Here, the class labels for the file name are regarded as real and fake. We receive 2820 instances of data and two class labels when we combine the two files.

5.1 Data Pre-processing:

The dataset has 34 properties in total, which is a large number of attributes. We are working to simplify and hone the fundamental qualities during this process. Identification of the individual or profile is done using the characteristics ID, Name, and screen name.

As a result, we only choose one of these three traits from the group here; we choose the ID in comparison to the other two attributes. Moreover, for profile identification, the properties statuses count, followers count, friends count, and favourites count are crucial.

The dataset also includes a listed count, which isn't very useful in our opinion, thus we minimize this property. Also, it's crucial to know how old a profile is, therefore the characteristic created at converts the date and time into the number of days. Since a social network profile will undoubtedly have a unique URL, the URL element is removed. We take into account time zone in comparison to the other two variables because the attributes lang, time zone, and location can all be concatenated into one.

5.2 Data Splitting:

For further usage in making decisions, the pre-processed data is used. As a result, previously arranged data is maintained in a local database to produce training and testing data. Here, the training set consists of 70% of the data instances that were chosen at random. The algorithms are tested using the 30% of the data that was randomly chosen. Furthermore, experimentation also makes use of the 80-20% ratio.

5.3 Training of Data and building model:

By using supervised algorithms like random forest and super vector machine (SVM) we have to train the data by using useful attributes and build a model using algorithms.

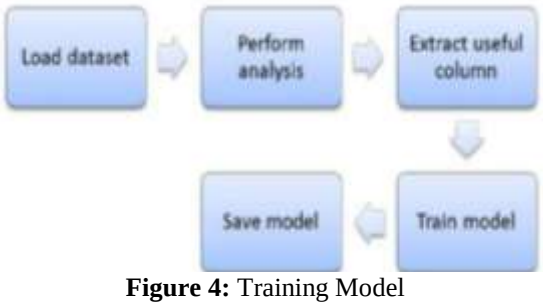

Figure 4: Training Model

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 1027-1032 www.ijaem.net ISSN: 2395-5252

5.4 Curve between the two Algorithms:

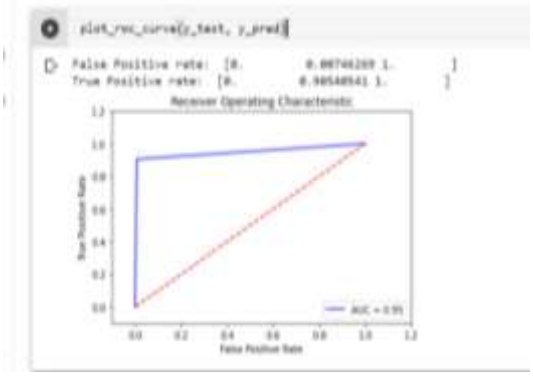

Figure 5: ROC Curve of RF Algorithm

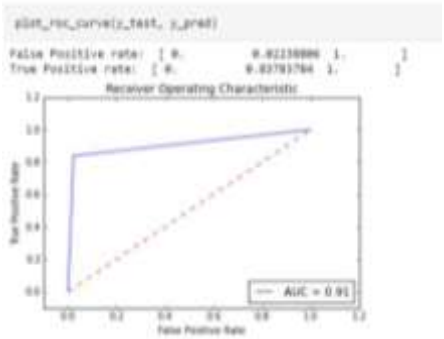

Figure 6: ROC Curve of SVM Algorithm

VI. FRONT END:

Figure 7: After entering genuine profile, it will give us output your account is real.

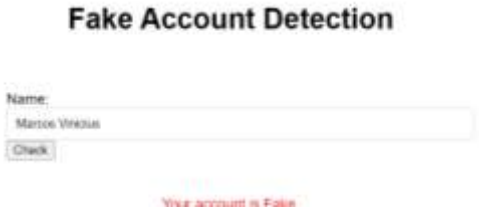

Figure 7: After entering fake profile, it will give us output your account is fake.

VII. RESULT AND ANALYSIS:

Metrics like precision, recall, and F1 score can be computed to assess the correctness of such systems. Recall represents the proportion of all false accounts that the system successfully recognised out of all those reported as fake, whereas precision indicates the percentage of accounts flagged as fake that are actually phoney. The F1 score gives a single statistic that balances both metrics and is the harmonic mean of precision and recall.

Based on this accuracy we found that the random forest approach outperforms the super vector machine after testing the two algorithms against the data. As a result, using the random forest method will provide us the highest level of accuracy for all account detection processes.

Random Forest	Super Vector Machine
0.9432624113475178	0.904255319149

CONCLUSION AND FUTURE SCOPE:

a. Conclusion

This project includes the implementation of a model that can identify fraudulent profiles on social media platforms. The model is user-friendly, allowing users to interact with it easily. It also lets users identify fake profiles or social media bots, protecting them from potentially harmful content. Attack on social media; we employed classification algorithms to identify false profiles; random forest has higher accuracy in huge datasets.

b. Future Scope:

Although we only utilized a tiny dataset from Kaggle to train the classifier, in the future we can use a larger dataset and design it separately for desktop applications and mobile applications, allowing users to use this system on a mobile device more ef ectively for spotting fake profiles.

REFERENCES:

[1]. Tiwari, V. (2017). Analysis and detection of fake profile over social network. 2017InternationalConferenceonComputing,CommunicationandAutomation(ICCCA).

[2]. (2018) Political advertising spending on Facebook between 2014 and 2018. Internetdraft. [Online]. Available: https://www.statista.com/statistics/891327/political-advertising-spending-facebook-by-sponsor-category/

[3]. (2012) CNBC. Facebook shares drop on news of fake accounts. Internet draft. [Online].Available: http ://www.cbc.ca/news/technology/facebook-shares-drop-on-news-of-fake-accounts-1.1177067

[4]. Khaled, S., El-Tazi, N., & Mokhtar, H. M. O. (2022). Detecting Fake Accounts on socialmedia.2022IEEEInternationalConferenceon Big Data(Big Data)

[5]. R. Nithin Reddy &Nitesh Kumar, "Automatic Detection of Fake Profiles in Online SocialNetworks,"ComputerScienceandEngineering,"National InstituteofTechnology,Rourkela.