Grenze International Journal of Engineering and Technology, June Issue

Anonymous Detect: A Machine Learning Model for Detecting Fake Profile on Instagram

Mukta Wagh[1], Vinay Sonwane[2], Harsh Kumbhare[3], Vijay Surwase[4], Ankit Zade[5] and Chinmay Tekade[6]
[1-6]Dept. of Information Technology, Yeshwantrao Chavan College of Engineering (YCCE), Nagpur, India
Emails: tekadechinmay10@gmail.com

***Abstract*— Fraudulent profiles are constant threats in the security landscape of social media, thus, the extent to which they affect privacy and trustworthiness of information is dire. Conventional detection algorithm, like the heuristic or fixed rules models, is becoming ineffective due to the manoeuvres of malicious parties. Our model is called AnonymousDetect, and it is a machine learning algorithm that can be used to detect fake Instagram accounts. The gist of the system is an optimized XGBoost classifier that is trained on a well-designed profile attributes. In order to solve the typical problem of high profile dataset imbalance between classes, we use the Synthetic Minority Over-sampling Technique (SMOTE) to create a balanced and bias-free training regime. The performance of the anonymousDetect was strictly assessed and the accuracy of the classification was 93.4**

***Index Terms*— Fake Profile Detection, Social Media Security, Machine Learning, XGBoost, Synthetic Minority Over-sampling Technique, Cybersecurity, Instagram.**

I. INTRODUCTION

On the one hand, social media has transformed the world communication process, but on the other hand, it has also revealed significant security gaps as a result of the expansion of false or fake profiles. Several parties with ill intentions regularly use such fraudulent accounts to support activities such as spreading false information, online fraud, unlawful gathering of human information, and influencing the masses [1], [2].

Since social platforms have grown fast and become complicated, it is no longer an option to rely on manual verification. This requires the creation of automated systems that can be dynamically adjusted to the ever-changing digital threats [3]. The early false profile detection methods were mostly focused on rules-based and blacklist filtering. This was never able to be flexible enough to respond to the ever-changing strategies of the bad users [4].

This deficiency led to the emergence of machine learning (ML) methods which are currently taking over the scene. The advantage of ML is that it can be used to detect complex and large-scale discriminative patterns automatically and thus the detection can be much more scalable and accurate [5]. We concentrated on gradi- ent boosting algorithms, mostly XGBoost, which has been repeatedly confirmed to effectively merge data of various types of features and model data relationships effectively [6].

The extreme lack of balance between classes naturally faces this issue, which is the difficulty of building strong classifiers since in the social media of reality, authentic accounts are many times fewer than fraudulent accounts. This bias often generates skew in the model, which translates to a large decrease in the accuracy of the minority class of fake.

Grenze ID: 01.GIJET.12.2.5_22

Anonymous Detect tries to evade this negative vulnerability with the help of the Synthetic Minority Oversampling Technique (SMOTE) [7]. SMOTE artificially increases the size of the minority group, thus forming a balanced set of data, which reduces training bias.
The main part of the work is the creation of a highly accurate, balanced, and interpretable model of detection, optimized to work with Instagram profiles. The unified framework combining advanced data preprocessing, feature engineering, class balancing and fine-grained model tuning attains a better performance than the traditional baselines. Most importantly, the system has a user-friendly interface that is real-time and enables end-users and the platform moderators to easily scrutinize suspicious profiles [8].

A. Applications

Social Media Platforms: AnonymousDetect allows increasing the level of user safety and the credibility of the content because it will automatically detect and flag inauthentic accounts on the largest social networks such as Instagram, Facebook, and Twitter [1], [2].
Online Dating Services: The platform proactively protects users against scams and fraudulent transactions in dating apps like Tinder and Bumble by properly detecting fraudulent dating websites [3].
E-commerce and Review Sites: AnonymousDetect ensures the credibility of product and seller profile comments in online stores such as Amazon and Yelp by helping to detect fraudulent reviewer accounts [4].
Cybersecurity and Law Enforcement: With the help of the insights provided by AnonymousDetect, security officials and investigators can better follow malicious groups that are staging organized misinformation, perpetrate online fraud, or harass others [5], [6].

II. LITERATURE SURVEY

Fake profile detection has received a wide body of research and led to a plethora of different methods this includes behavioral analysis, social graph mining, and classification using machine learning. The main goal of the early detection was the attempt to utilize the user activity and behavioral patterns.
The analysis of such a platform as Twitter effectively used the following measures: the frequency of posts, the ratio between followers and those followed, and signs of content duplication as effective tool to distinguish between the human-controlled and the automated ones [1]. Other attempts along this behavioral focus made use of traditional classifiers like Support Vector Machines (SVM) and Naive Bayes to classify ac- counts (e.g. humans, bots or cyborgs) with reference to time relevant activity and engagement patterns [3].
A secondary, critical one incorporated network structure and content-based signals. The study emphasized the significance of social graph features and the content of this message between users and employed relationship patterns and frequency of communication to discriminate spammers [2]. Additional progresses included the analysis of new kinds of content attributes, including the proportion of outgoing URLs, repetitive patterns of hashtags, and the inconsistency of posting times, which could better detect them [4].
Taking into consideration the growing sophistication of malicious techniques, there have been some attempts at combining graph-based and textual analysis functions in hybrid models that enhance precision in various social media platforms [6]. Moreover, Varol et al. [5] indicated that model interpretability is essential and provides transparency in prediction, and this is essential in sensitive fields, where the user trust and privacy are central.
This is a general survey that shows that there is a wide variety of methods available, but there are still considerable research issues. The area still has to deal with the technical challenges of extreme data imbalance and requirement to have interpretable and real-time capable models. The project AnonymousDetect is based on these already existing foundations, where an optimized XGBoost classifier is used with SMOTE-balanced training, which is specifically configured to the peculiarities of the Instagram profile.

III. PROPOSED METHODOLOGY

AnonymousDetect system has been designed to be a powerful, pipeline-based detection system. This methodology involved the data collection process, data cleaning, feature engineering through sophisticated engineering, training models supervision, comprehensive performance analysis and finally deployed as a real-time web application.

A. Data Acquisition and Preprocessing

The data utilized in the current study was collected in an open-source Instagram profile depository on Kaggle. The sample used contained 3000 labelled user accounts, which were categorised as fake and genuine. Notably, in the former dataset, the imbalance of classes was high because it consisted of only.

It has 400 counterfeit profiles and 2,600 real accounts. This skewed nature became one of the primary threats of challenging the final classifier to that of majority class, hence restricting the capability of the classifier to make predictions of minority classes.

Synthetic Minority Over-sampling Technique (SMOTE) had been employed in order to solve this imbalanced representation. The SMOTE created synthetic examples of the underrepresented class, and could achieve a complete balanced training set of 3,000 fake and 3,000 real profiles. Data preparation was performed in two basic steps, in order to ensure data integrity and immunity to outliers, the gap in feature entries was imputed with a median-based imputation. Min-Max normalization was performed after imputation and it normalizes all possible numeric variables to a range of 0-1. This critical scaling process makes sure that the features that end up being large by definition never dominate the objective function of the model during the learning process.

B. Feature Engineering and Real-Time Prediction

Only profile-level and behavioral measures that are very discriminative were automatically harvested and profile attributes were extracted by the Instaloader API. The key features that are designed to undertake the classification task are: •Follower to Following Ratio: This is a quantitative data of the ratio between the level of popularity (followers minus) of a particular account and its outbound follow policy. • Privacy and Verification Status: A Binary variable of key profile authentication descriptors. • Bio URLs: The amount and the exis- tence of external links that are given on the biographical text of the account. • Post count and frequency: It is a historical activity of the user measure that demonstrates the degree of involvement or activity of the use. Normalization and combination of these attributes will lead to the final input feature vector which is to be used in the classification of profile authenticity in the real time application.

C. Model Selection and Training

XGBoost (eXtreme Gradient Boosting) algorithm was selected due to its efficiency, high scalability and the ability to deal with complex and heterogenous tabular data effectively to classify data in situations of classification. A balanced dataset was strategically split in advance of the training procedure and 70 per cent of it was utilized in the training process and 15 per cent in the validation and the other 15 per cent in the final testing. The optimization of hyperparameters (such as the maximum tree depth, learning rate (), L1 L2 regularization ((,) was one of the most important ones that we made. Such fine-tuning was crucial to successfully manage the complexity of the model so that the risk of overfitting would be reduced and the ability of it to generalize on previously unknown data would be optimized.

IV. RESULTS AND OUTPUTS

A. Model Performance

Fig. 1 is the normalized confusion matrix and the key classification measures of the XGBoost model. The overall accuracy of the classifier is high (93.4) and especially in the case of fake accounts (98.3) and an effective performance in the cases of genuine accounts (89.6). The confusion matrix with high values in the diagonal visually justifies the low misclassification rates in the system. The classification report gives further details: the precision and the recall are high in both classes and the F1- score give an example of a balanced trade-off between sensitivity and specificity.

The Area Under the ROC Curve (AUC) of the model is 0.94, which once again confirms that the model can be used despite its discriminative power on difficult, imbalanced data. Indeed, such measures clearly show that InvisiDetect can be effectively used in the real-world to reduce any false positives and false negatives in order.

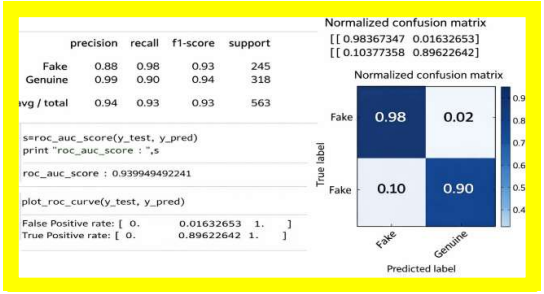

Figure 1: Normalized Confusion Matrix and Classification Report of the XGBoost model of the test set, which show high accuracy and distinguish between fake and legitimate accounts.to enhance platform security.

B. Application Interface Examples

The next figures denote the end-user perspective of the Streamlit application. Both of them show how the results of predictions and the features that support these are visualized in real-time situations. InvisiDetect marks a suspicious Instagram account as a Fake as shown in Fig. 2. The UI does not only present the classification output but also actionable options including the follower / following ratio, the account status, the number of posts and the recent activity measures. Through this provision of relevant statistics, this interface provides confidence in the recommendation of the system and transparency of the system to the user or moderator when making a decision.

On the other hand, Fig. 3 shows the UI when a real user profile is examined. In this case, the interface shows a clear distinction between detectives states with the account being marked as Genuine and supporting information being displayed including the profile verification status and the frequency of posting. This systematic presentation assures non-fake users are not falsely noticed so that user experience is not negatively affected and intervention is minimized.

To test the power of the model further, another instance of fake account detection is presented in Fig. 4 that ensures the flexibility of the model in the face of various adversarial behaviour and manipulative behaviour patterns. The supporting attributes remain active despite the input variability and prove the model to be ready to work with the reality of the work and to be audited by administrators.

Finally, Fig. 5 also supports the precision of the InvisiDetect on genuine accounts, which proves that the solution can scale and prosper within a large spectrum of user profiles without the needless impediments, which are essential to keep the social network users trustful.

C. Model Comparison: XGBoost vs. Random Forest

In order to measure the progress of the previous work, a systematic comparison with the previous Random Forest is conducted. All the major metrics are provided in Table I. Although the initial results were good in Random Forest, it is seen that XGBoost has a slightly better accuracy and a more consistent outcome especially with hard cases. The two models both have an AUC of 0.94, although XGBoost has always indicated better precision and recall rates in the 'Fake' class.

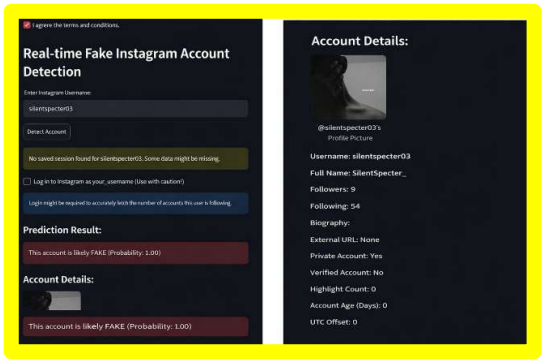

Figure 2: UI display to identify a fake Instagram account with a critical supporting information and profile characteristics

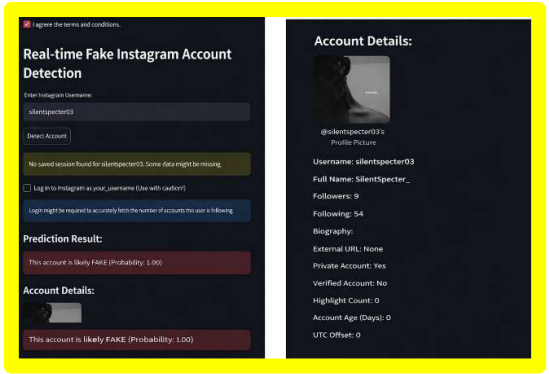

Figure 3: UI output of determining a true Instagram profile, visualizing authenticated status, feature level

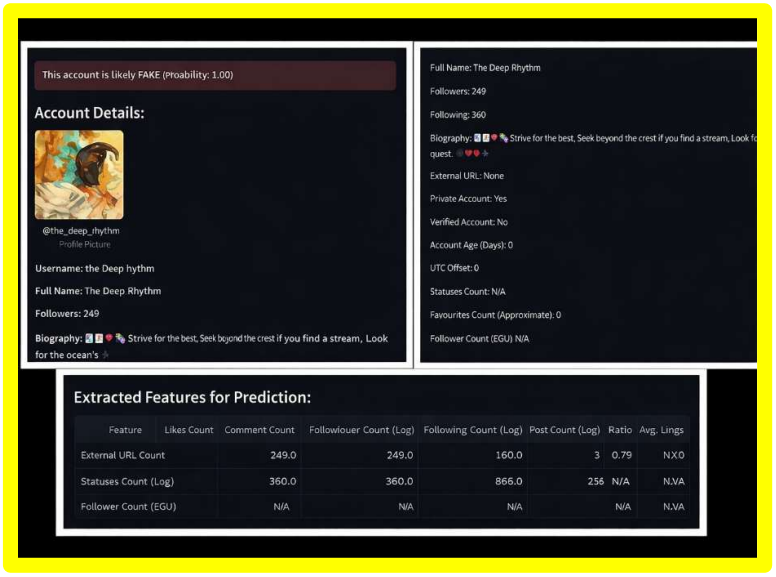

Figure 4: Alternate of fake profile detection repeatability, which supports model generalizability and robustness

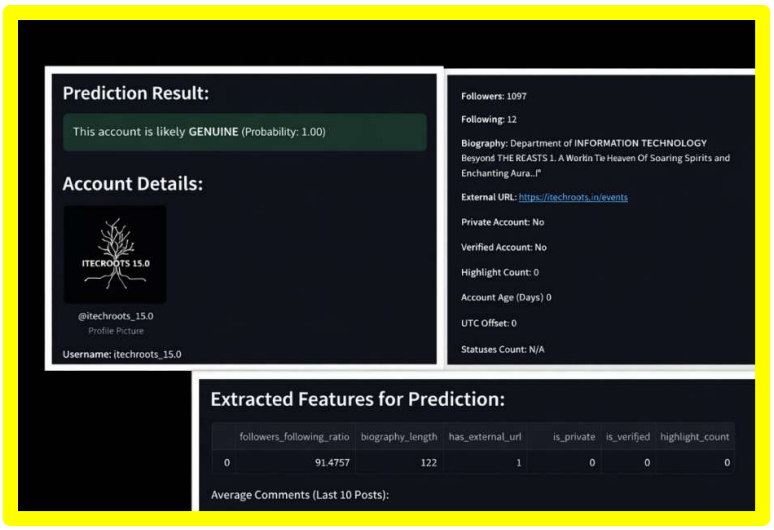

Figure 5: Demonstration of genuine user detection, reinforcing model reliability and trustworthiness for everyday use

TABLE I: COMPARISON OF PERFORMANCE: RANDOM FOREST VS. XGBOOST

Metric	Random Forest	XGBoost
Accuracy	0.94	0.934
Precision (Fake)	0.90	0.91
Precision (Genuine)	0.99	0.99
Recall (Fake)	0.99	0.98
Recall (Genuine)	0.90	0.89
F1-Score (avg)	0.94	0.93
AUC	0.94	0.94

Table I shows the further improvement of model balance with the use of transition to ensemble boosting methods, particularly when data imbalance occurred. Fig. 6 shows the learning and validation scores of the Random Forest where a greater variance and moderate overfitting is seen in the difference between the training and cross- validation scores. This comparison highlights the reason why the transition to XGBoost resulted in more generalization and an improved result on difficult cases.

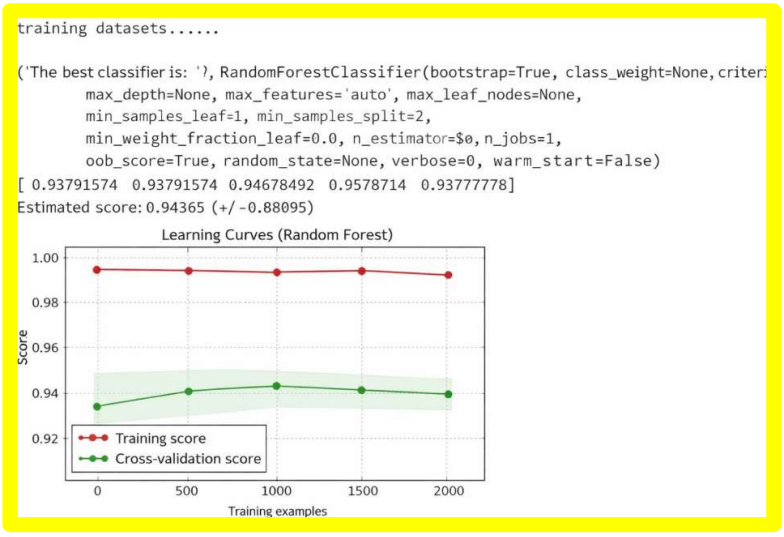

Figure. 6: Learning curves in the case of a Random Forest model show higher variance and generalization differences in validation compared to the XGBoost.

In Fig. 7, the normalized confusion matrix for Random Forest further highlights its slightly reduced performance in accurately classifying genuine users—while still offering strong results, the improvement via XGBoost is clear when reviewing end-user impact and the likelihood of real-world misclassification.

Figure 7: Normalized confusion matrix for Random Forest, showing strong but slightly less balanced results than XGBoost, especially for the "Genuine" user class.

IV. Conclusion

The InvisiDetect relies on state-of-the-art machine learning to detect counterfeit social media accounts effectively to enhance the integrity of the platform and the safety of the user. Our model addresses the imbalance in data performance by employing the SMOTE method and shows a high accuracy and the AUC of XGBoost. This solution can help moderators and users stop scammers and improve the social media ecosystem by helping them forecast fraudulent accounts in real-time and with a simple interface.

Acknowledgment

The authors gratefully acknowledge the support and re- sources provided by the Department of Information Technol- ogy at Yeshwantrao Chavan College of Engineering, Nagpur.

REFERENCES

[1] Yang, C., Harkreader, R., & Gu, G., "Empirical evaluation and new design for fighting evolving Twitter spammers," IEEE Trans. on Information Forensics and Security, 2011.

[2] Stringhini, G., Kruegel, C., & Vigna, G., "Detecting spammers on social networks," Proc. ACM Confer- ence on Computer and Communications Security, 2010.

[3] Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S., "Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?," IEEE Trans. Dependable and Secure Computing, 2012.

[4] Cao, Q., & Caverlee, J., "A large-scale analysis of fake followers on Twitter," Proc. ACM Web Search and Data Mining, 2014.

[5] Varol, O., Ferrara, E., Davis, C., Menczer, F., & Flammini, A., "Online human-bot interactions: Detection, estimation, and characterization," Proc. Int. AAAI Conf. on Web and Social Media, 2017.

[6] Patel, K., & Singh, M., "Hybrid machine learning model for fake profile detection in online social net- works," Proc. Int. Conf. Data Science and Machine Learning, 2020.

[7] Shah, M., Joshi, H., & Gupta, N., "Implementation of social media fake profile identification using a novel machine learning technique," in Proc. IEEE ICTBIG, 2023.

[8] Kumar, R., Srinivasan, S., "Detection of fake social media profiles using supervised learning," Int. J. Comput. Sci. Inf. Secur., 2022.