Fake Profile Detection Using Essential Machine Learning Models

Shahadat Hossain Sagor
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
shahadat.hossain.sagor@g.bracu.ac.bd

Israt Kayesh Ipsit
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
israt.kayesh.ipsit@g.bracu.ac.bd

I. Introduction

Due to technological advancement the dependency on online social media apps such as -Facebook, Instagram, twitter etc. has increased a lot in different sectors like- marketing, communication and personal interaction. This reliance has lifted a significant number of security and privacy concerns with the expansion of creating fake accounts on these platforms. Identity frauds, fake profiles and different bot accounts are spreading fake news, deceiving innocent people and also making a lot of celebrity and non-celebrity people the prey of cyber bullying. The first and foremost thing to eradicate these malicious activities is detecting these fraudulent accounts. Nowadays these accounts are detected using Machine Learning algorithms. Most widely used ML algorithms for this regard are- Random Forest, Support Vector Machine, Naïve Bayes and Deep Learning techniques. Studies have also suggested innovative and advanced methods like Sybil Rank, Integro systems, sentiment analysis and entropy minimization which examine the attributes and behaviors of the user leading to increase the rate of detection accuracy and eradicate the effect of fake accounts. The comprehensive aim of the researchers is to create robust and dynamic models which will preserve the integrity of the social media networks and more capable of differentiating between fake and original profiles.

II. Literature Review

The study [1] mainly emphasises on detecting whether the accounts in social media are operated by bots or humans. TwiBot-20 dataset has been used for research purposes. This dataset contained id, account description, language, screen name, verified or not and the target variable which is the type of account (used by bot or human). Before implementing the machine learning model the data was preprocessed by removing the null values then converting the categorical values into numerical values using TF-IDF vectorizer and at last correlation between the features were checked to avoid redundancy. Almost 300 features were present in this dataset. Machine learning models work better and give more accuracy in preprocessed data. Most widely used ML models were used in this dataset like- Random Forest Classifier, Support Vector Machine and combining both Random Forest and Naive Bayes an ensemble learning algorithm was used. N-estimators hyperparameters were used in RF classifiers and kernels were used in SVM which lead to maximize the margin and minimize the error. At the end of the study Random Forest gave the highest accuracy among all the models which was 87% and ensemble learning gave the second highest accuracy of 86%. But SVM gave an average accuracy of 71% as the dataset had high dimension and lacked effective feature scaling. Not only this but also the computational intensity of SVM inhibited the performance. The Matthews Correlation Coefficient(MCC) of the random forest classifier was 0.7048. The accuracy rate of this study has successfully surpassed the findings of previous studies. Though the study emerged successfully in higher accuracy rate it has some drawbacks. Firstly, the TF-IDF vectorizer was effective but it limited the further exploration. Secondly, in the case of SVM the hyperparameter tuning has hindered the performance of the model. Lastly, the dataset was quite limited which lessened the further analysing of account types. Despite its high accuracy it has a lot of scopes of improvement. Advanced techniques like Word2Vec or BERT could work better and be more fruitful in case of advanced test representation techniques. Grid or random search type techniques could have been used for further optimization of models. Expanding the features of the dataset could improve the generalization of the model. Moreover, techniques like Stacking or Boosting can increase the predictive accuracy. Incorporating network-based features, such as the relationships and interactions between accounts, could also provide further insights. While the Random Forest model performed admirably, achieving a high level of accuracy, there is significant scope for enhancing model performance through better feature selection, model tuning, and the use of more advanced machine learning techniques.

The research aims to detect extremist users and fake accounts on social media, particularly Twitter, using machine learning techniques. In this study [2] no explicit dataset has been used, rather a large dataset of millions of users which included the accounts which were suspended due to extremist activities. The study also analyzed tweets from almost 25000 regular users involved in extremist content. It is focused on three factors- detection of the extremist users, estimation of regular users adopting extremist content and prediction of

users engaging with extremists. Different Machine Learning models have been compared in this research such as- Support Vector Machine and Ensemble Classifiers. Models like- Adaboost, Reinforcement Learning model and SVM were used in this research. Noise removal and data normalization were used which resulted in the improved dataset quality, with algorithms inspired by natural processes such as Artificial Bee Colony and Ant Colony Optimization applied to discard non-significant attributes. Ensemble Classifiers also improved the prediction accuracy. While comparing the models reinforcement learning model achieved the highest accuracy of 87.11% and 49.75% F1-score and 49.90% PR-AUC, Adaboost with second highest accuracy which is 85.91% and 47.54% F1-score and 49.53% PR-AUC and SVM with average accuracy 68.05% and 32.16% F1-score and 27.76% PR-AUC. However, the study notes that dealing with imbalanced datasets (where fake accounts are outnumbered by real accounts) remains a challenge, as fake profiles are hard to detect due to class imbalances. Several drawbacks can be seen due to limited generalization throughout the social networks. Relying on content features had the tendency to allow extremist and fake accounts to dodge the fake account detection as the contents can be easily altered. Though there are drawbacks it can be easily avoided by implementing some advanced techniques like SMOTE, hybrid models using both supervised and unsupervised learning and semi-supervised learning. Also, if explainable AI can be introduced then it will improve the system transparency. Integrated user interaction patterns can also enhance the accuracy rate. To sum up, the research has been depicted that ML models have shown significant performance in detecting fake accounts.

In recent times the studies use the datasets from different social media platforms and these datasets range from 700 to 82,000 instances. The Instagram dataset obtained from Kaggle was used for this fake profile detection paper [3]. Eleven attributes were present in the dataset and a total of 700 instances were there. 80% of the data were used for training and 20% for testing. Renowned ML models SVM, K-Nearest Neighbour, RF, Multi-Layer Perception, Naive Bayes Classifier and Logistic Regression were used in this research. Performance of the models were based on some criterias such as accuracy, F1-score, recall and precision. Random Forest had shown the best performance with 96% accuracy and precision of 95%. SVM and Logistic Regression had also shown better performances with 94% and 93% accuracy and precision respectively. Multilayer and K-Nearest Neighbour had the same accuracy of 91% but precision 91% and 86% respectively. Naive Bayes had shown the least accuracy rate and precision of 74% and 82% respectively. One of the main drawbacks of this successful research is the risk of overfitting with models like Random Forest and SVM if they are implemented in small datasets. MLP based models face struggles with linguistic complexities such as cultural and sarcasm references. Also the computation of real time detection is comparatively more expensive as complex feature engineering and large scale ensemble models are used. Though there are drawbacks it can be also improved through using more advanced techniques of Deep Learning such as CNNs and RNNS. CNN and RNN are more applicable for accurate image and text analysis. It uses edge computing or cloud computing for optimising the real time detection. Hybrid models that integrate multiple Machine Learning techniques or combine Machine Learning with rule-based systems may enhance accuracy and efficiency. Though this study has shown promising results for detecting fake accounts but by using the above mentioned advanced techniques and comprehensive datasets significant amount of improvement can be made in this field.

Detecting fake accounts and bots on social media platforms such as Facebook, Twitter, and Instagram has become a crucial area of research due to the rise in malicious activities like spamming and fraudulent behaviour. Success in this area heavily depends on the quality and diversity of datasets used in these studies. This research relied on datasets collected through APIs, user profiles, and interactions from social media. Datasets from Facebook contain both types of accounts genuine or fake with the behavioural data. The study [4] combines data from multiple platforms like Facebook, LinkedIn, and Twitter to provide a generalized analysis of fake accounts. After collecting the datasets they are preprocessed for better results after applying different ML models. Data preprocessing plays a critical role, where issues like missing values, outliers, and inconsistencies in the dataset are addressed. Models like Neural Network, SVM and Random Forest Classification Technique were used in this research. SVMS handled the high dimensional data very well compared to other models. Random Forest Classifier, a kind of ensemble learning algorithm, shows high accuracy especially when working with large datasets and also reduces the variance and overfitting problems faced in other models. After implementing the models it was seen that the RF classifier achieved the highest accuracy of 94.1%. It outperformed both neural networks and SVM performance which was 81% accuracy rate. The main drawback of this research is the high computational cost for using the neural network algorithms. SVMs are quite prone to overfitting which hinders its overall performance and it is also inefficient with large datasets. Also, SVM has dependency on Kernel and the wrong choice of kernel leads to a suboptimal result. Although random forests reduce overfitting compared to individual decision trees, shallow trees can still underperform, especially with more complex datasets. This research has a lot of space for improvement. In case of Neural Networks using dropout, L2 regularization can provide the model to increase the generalization performance. To improve SVM's scalability, future work could focus on exploring more efficient algorithms like stochastic gradient descent (SGD) or applying kernel approximation techniques to reduce computational complexity. Using hybrid kernels can increase the performance of SVM in complex datasets. Regularization techniques should be more optimised for preventing the overfitting in high dimensional datasets. In conclusion, while the random forest classifier demonstrated superior accuracy in this study, optimizing neu-

ral networks and SVMs through advanced techniques offers promising avenues for further improving classification performance and model efficiency.

The study [5] focuses on detecting fake Instagram accounts using the Instagram Fake Spammer Genuine Accounts dataset from Kaggle. Total 696 instances were used here among which 576 were for training purposes and 120 for the testing of models. Firstly, the data were preprocessed and then data cleaning for better results. Distribution graph and correlation matrix is used for preprocessing. The study implements multiple Machine Learning models, including Random Forest (RF), Artificial Neural Network (ANN), Logistic Regression (LR), Bernoulli Naive Bayes (BNB), and Support Vector Machine (SVM). The research was conducted through two experiments. In the first experiment, a default model was used which was imported from scikit learn to observe its performance on the following dataset. Default parameters were used in all models. The Random Forest model obtained a high accuracy rate of 92%. Following that logistic regression had an accuracy rate of 91%. SVM had an accuracy rate of 89% and BNB yielded an accuracy rate of 87%. ANN had shown least accuracy of 82%. After obtaining high accuracy in the 1st experiment a new feature where followers are more than following or not was added in the second experiment for comparison. After tuning the models with this feature the accuracy rate in all the models increased except SVM. Both Random Forest and Logistic Regression had an accuracy rate of 93%. From giving least accuracy in 1st experiment ANN had shown 2nd highest accuracy in second experiment of 92%. The accuracy rate of SVM had fallen and was given the lowest rate of 88% because of dimensionality. Despite these successes, several limitations were identified. Small amount of data had limited the potential of the model. Due to high dimensionality SVM yielded worse performance in the second experiment as there was a need for more data of new features. The feature selection was somewhat limited in the above study. But improving data quantity and diversity and also including other social media platforms will increase the accuracy of the models. Feature Engineering could also be expanded by analyzing user bios, posts, and employing advanced text and image processing techniques. CNN and RNN should be introduced in this study for improving the accuracy and for precisely analysing the images and text based features of the profiles. Grid or random search would optimise the models and will provide better results. Data or cloud based architectures should be used so that models can be blended into the real time detection system that would be more effective in identifying the fake profiles. Through further optimization and data expansion the study will provide a higher accuracy rate.

The detection of fake accounts on social media platforms has become a critical area of research, with various studies exploring Machine Learning and statistical approaches to tackle this issue. Datasets used in the study [6] are drawn from platforms like Facebook, Twitter, and Instagram, capturing user activities such as posts, comments, friends, and profile attributes to identify fake accounts. The study had used large datasets collected over time, including profile attributes and user behaviours, to enhance the detection of bots and fake accounts. The Supervised Machine Learning models used in this study were K-Means Clustering, Naive Bayes, SVM and Adaptive Boosting. For detecting the complicated patterns in fake profiles neural networks were being used. Image processing techniques were used for image recognition. Different types of approaches were being used to minimize the false positives and negatives while detecting fake profiles. This study has shown prominent results using the above models. K-Means Clustering has shown 75.3% of accuracy rate. SVM models reached 95.8% accuracy, while Gaussian SVM achieved up to 97.6% in detecting fake accounts. Naive Bayes demonstrated a 92% detection rate but had slightly lower overall accuracy. However, certain limitations persist. SVM models are quite prone to overfitting. Naive Bayes shows better performance but when datasets get more complex its performance slows down while distinguishing more complex behaviour. Future research could focus on improving dataset balance with advanced resampling methods and developing real-time detection models for dynamic social networks. Moreover, using Natural Language Processing (NLP) techniques can analyze the contents of users more deeply. Improved image detection algorithms could help identify fake accounts using manipulated images, further enhancing detection accuracy. Though the study is giving better results in detecting fake accounts, it can be evolved more by using advanced techniques mentioned above.

III. Conclusion

With increasing usage of social media the number of fake accounts are also increasing in a significant amount causing lots of problems and harassment to people. Detecting fake accounts on social media networks has become a prominent space for research using different Machine Learning models and statistical methods. The studies reviewed highlight the use of diverse datasets, ranging from user activities to intricate behavioural patterns, which significantly influence the fruitfulness of detection models. While algorithms like Random Forest, Support Vector Machine, and Naive Bayes have shown promising results, achieving high accuracy rates, challenges such as data imbalance, overfitting, and the evolving tactics of malicious actors persist. If we were to do this study we would introduce the hybrid models combining Random Forest with XGBoost and Deep Learning where CNN would be used for image and RNN for text analysis. It would increase the accuracy significantly. We would also use GNNs which would analyse user relationships and interactions alongside XAI providing transparency and making the detection process more reliable. Using Word2Vec or BERT for feature engineering would improve the results. Though there are limitations in the above studies, embracing the above mentioned methods or models will enhance the accuracy rate of detecting the fake accounts and will create safe space for the users in the social media world.

REFERENCES

[1] Umbrani, K., Shah, D., Pile, A., and Jain, A. (2024, January). Fake Profile Detection Using Machine Learning. In 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS) (pp. 966-973). IEEE.

[2] Patel, K., Agrahari, S., and Srivastava, S. (2020, June). Survey on fake profile detection on social sites by using machine learning algorithm. In 2020 8th international conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO) (pp. 1236-1240). IEEE.

[3] Joseph, J., and Vineetha, S. (2023, November). Fake Profile Detection in Online Social Networks Using Machine Learning Models. In 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE) (pp. 1-5). IEEE.

[4] Kathiravan, M., Parvez, S. J., Dheepthi, R., Jayanthi, R., Gowsalya, S., and Sekhar, R. V. (2023, January). Analysis and detection of fake profile over social media using machine learning techniques. In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1164-1169). IEEE.

[5] Ekosputra, M. J., Susanto, A., Haryanto, F., and Suhartono, D. (2021, December). Supervised machine learning algorithms to detect instagram fake accounts. In 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 396-400). IEEE.

[6] Sen, S., Islam, M. I., Azim, S. S., Norin, F. A., and Shuha, S. T. (2021). Fake profile detection in social media using image processing and machine learning (Doctoral dissertation, Brac University).