International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Feature-Engineering Approach to Machine Learning for Detecting Fake and Bot Accounts on Social Media

Abhishek Agarwal[1], Dr. Shivangi Barola[2]

[1]Research Scholar, Faculty of Computer Science, Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan
[2]Assistant Professor, Research Scholar, Faculty of Computer Science, Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan

Abstract
This paper introduces a machine learning-based classifier to identify fake accounts, bots and malevolent profiles on social networks such as X (Twitter), Facebook, and Instagram. The study combines behavioural, linguistic and network-based features to determine the genuine and automated users using the Random Forest, Support Vector Machine (SVM) and Neural Network models. As per the experimental findings, Random Forest was found to be the most balanced in terms of accuracy and interpretability with a detection accuracy of over 90 percent, whereas the Neural Networks offered more recall of complex cases. The analysis of explainability based on SHAP and LIME proved the superiority of behavioural and network predictors. The paper provides a scalable, explainable, ethically-grounded method of enhancing digital trust and platform security.

Keywords: Machine learning, malicious profile, bot detection, Linguistic Features, Network Analysis, Random Forest, Support Vector Machine (SVM), Feature Engineering, Behavioural Analysis

1. Introduction

Facebook, Instagram, X (twitter) and WhatsApp are some popular social media platforms that are used in digital communication and digital information exchange. As a wide range of people use these social media platforms the most concerning issue is privacy and security. Their exposure and size have led them being the most targeted objects of misuse by false accounts, robots and malicious profiles. These fake accounts in social media platforms are used to misrepresent narratives on the internet by enhancing fake news and propaganda. On the other hand, bot accounts are used to manipulate algorithms and recommender systems or promote political speech (Tunç *et al.* 2022). Moreover these bots and fake accounts are used for cybercrimes such as phishing and data breaches. These accounts can influence in times of elections, geographical conflicts and disrupt social trust. The traditional detection methods, mostly rule-based or keyword based, are facing challenges to keep up with the complexity of these anomalies (Hayawi *et al.* 2023). These fake accounts and bots can now imitate the legitimate user behaviour to escape the detection. Machine learning methods provide more dynamic and scalable alternatives as this method can learn the pattern of the data to identify the profiles according to their behavioural, language and network features (Sarker *et al.* 2021). By using machine learning and data

analytics future threats can be predicted and actions can be taken according to it. The proposed study aims to develop and test machine learning that combines advanced features engineering methods to detect the fake accounts and bots.

1.1 Objectives

The objectives of the study are:

- To improve the detection accuracy and reduce the number of fake accounts and bots.
- To Design a deployable and realtime detection framework using algorithms like neural networks, SVM and random forest.
- To enhance detection accuracy through feature engineering.
- To develop a scalable real-time system capable of detecting fake and automated accounts on live social media streams.

2. Literature review

2.1 Early Approaches and Limitations for Detecting Animalities

Prior studies of social media fraud detection were mostly based on rule-based approaches, with certain thresholds or patterns. The commonly used detection system uses a trained model that uses labelled data to draw out user profiles, network structure, or user activity to analyse the fake accounts or bots (Mohammed *et al.* 2024). Though these methods were computationally effective, it was becoming ineffective as malicious users started modelling the normal human behaviour. In addition, these could not be as flexible in responding to dynamic and diverse platform environments.

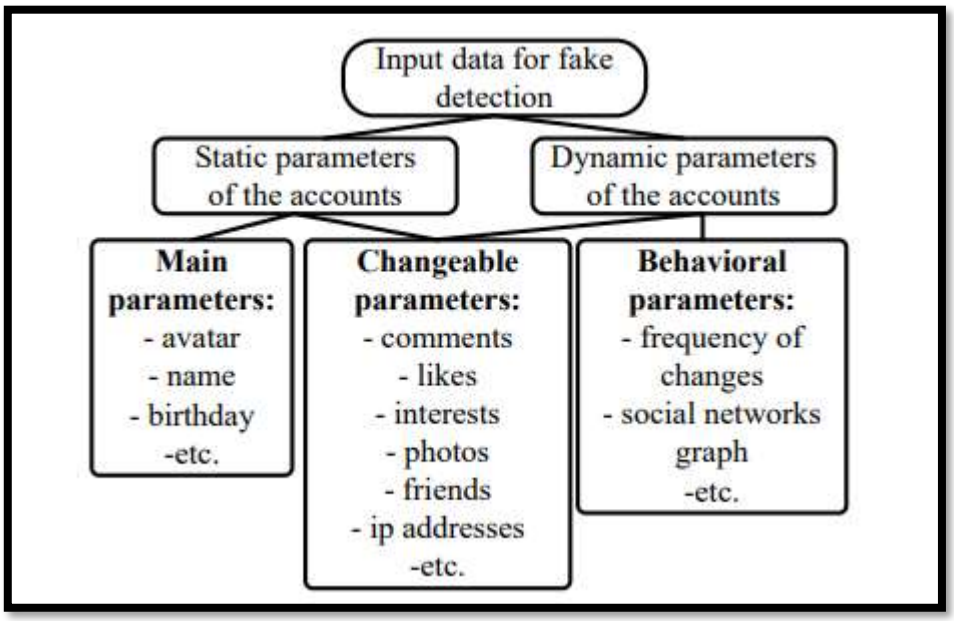


Figure 1: The input parameters for fake detection

(Source: Voitovych et al. 2022)

2.2 Emergence of Machine Learning in Fake Account Detection

As social media is vastly used the security for it is crucial and machine learning can enhance the detection of things easily. Researchers used managed machine learning algorithms such as random forest, decision trees, and support vector machines to categorise accounts using behavioral and network-based characteristics (Voitovych *et al.* 2022). Random Forests were especially useful when it comes to working with non-linear associations among variables and high dimension feature spaces. SVMs,

however, were able to perform well with small to medium sized data sets, and it has got high generalization ability (Ioannou *et al*. 2021). Although there were such developments, data imbalance and model interpretability were still persistent problems.

2.3 Feature Engineering as a Core Challenge

Detection of the fake accounts and bots in social media platforms is a complex problem and a critical aspect that has affected performance in models was feature-engineering (LAM *et al.* 2025). Some of the behavioral indicators, including retweet ratios, posting intervals, and the variability of engagement, typically depict non-human behavior which algorithms can take advantage of. Similarly, such linguistic indicators as excessive use of hashtags, text repetition, or text extremity are signs of automation or simultaneous manipulation.

2.4 Deep Learning and Representation Learning

Deep learning models in recent years have radically changed detection of fake accounts by the capability to learn non-linear feature representations in a complex way automatically. Time series Neural Networks (RNNs), Long Short-Term Memory (LSTM) have been used to learn temporal dynamics of user activity (Ghislieri *et al.* 2021). Graph Neural Networks (GNNs) and Graph Convolutional Networks (GCNs) provided even more expansive versions of this paradigm, by fulfilling the relationships between the users as graphs, including inter-user dependence and synchronized behavior (Zhou *et al.* 2022). Although these models perform better than traditional machine learning, they require significant computational resources and are ethically questionable in terms of explainability and privacy.

2.5 Metrics and Model Benchmarking in Evaluation.

Measures that are commonly used by performance evaluation in fake account detection are accuracy, precision, recall, F1-score, ROC curve, confusion matrix and Area Under the Curve (AUC) (Shah *et al.* 2024). It is crucial to balance precision and recall because excessive false negatives will leave malicious actors unpunished (Terumalasetti *et al.* 2025). The F1-score provides a balanced single value representation and it is useful when both false positives and false negatives carry significant consequences.

3. Research methodology

3.1 Research Design

The research design of the study is assumed to be quantitative and experimental with focus on the empirical testing, statistical validation and the development of the model. The study uses an organised way to incorporate data collection, pre-processing, feature engineering, modelbuilding, and performance validation. In contrast this study does not rely on theoretical knowledge only, but also uses experimental data such on both large scale datasets and live social media streams to ensure practical relevance, robustness and scalability. The systematic review of various machine learning algorithms, such as Neural Networks (NN), Random Forest (RF) and Support Vector Machines (SVM) under both control and real-life conditions are the focus of the experimental design. All the algorithms will be trained further to determine their effectiveness in differentiating between real and fake accounts and bot profiles. In addition, to minimise overfitting and improve the generalizability, cross-validation methods are used.

3.2 Study Area and Population

The research explores three popular social media networks, including Facebook, X, and Instagram. All of these together means that these platforms have billions of active users. Moreover, these platforms have a highly diverse mixture of interaction patterns, content types and engagement behaviours. The

target population will consist of confirmed users, legitimate accounts, potential fake accounts and bots. The heterogeneity of the population also makes sure that the models are being trained against at a realistic range of behaviours and mitigates the chances. By incorporating various user profiles, the research will formulate detection models which are resistant to various accounts and behaviour patterns.

3.3 Sample Design

The stratified random sampling strategy will be used whereby the dataset used will reflect the percentage of true users, fake profiles, and bots. Stratification is required because the population of social media is inherently unbalanced, with actual users by far exceeding the individuals with ill intentions. The training data will be based on the publicly available datasets, such as "Botometer Dataset" and "Fake Profile Detection Dataset" on Kaggle. As a manner of testing model performance in a live environment, more datasets will be acquired with platform-specific APIs, such as Twitter API and Facebook Graph API. The final data will consist of about 100,000 profiles, which is large enough to be statistically significant but not too large to be used in repetitive model training.

3.4 Data Collection

The method of careful sampling combined with extensive data collection and strict preprocessing conditions allow making the dataset robust, diverse, and representative. This is the base on which machine learning models with high accuracy, low false and real-time deployment on various social media platforms will be developed. When collecting the data, the strong compliance with the data privacy regulations and those intrinsic to the platform will be observed to guarantee ethical conduct. All information that can be identified to a person will be anonymized and data will be safely kept to analyse.

- Account metadata: Contains some basic profile information like data account creation, profile completeness, verification status and characteristics of a username.
- Network features: It shows the position and the connectivity of the account in the social graph. This comprises the following ratios, network centrality measures and reciprocity of interactions.
- Behavioural features: Patterns of account activity over time such as frequency of posting, time between posting, retweet/reply ratios, patterns of burst activity, variability of engagement, and stability of interaction.
- Linguistic features: The main variables will be sentiment polarity, keyword or phrase repetition, the use of hashtags, the frequency of emojis, the use of spam texts, and the diversity in the content. The characteristics are used to identify accounts that adopt monotonous, automatic, or spamming language to attract or steal users.

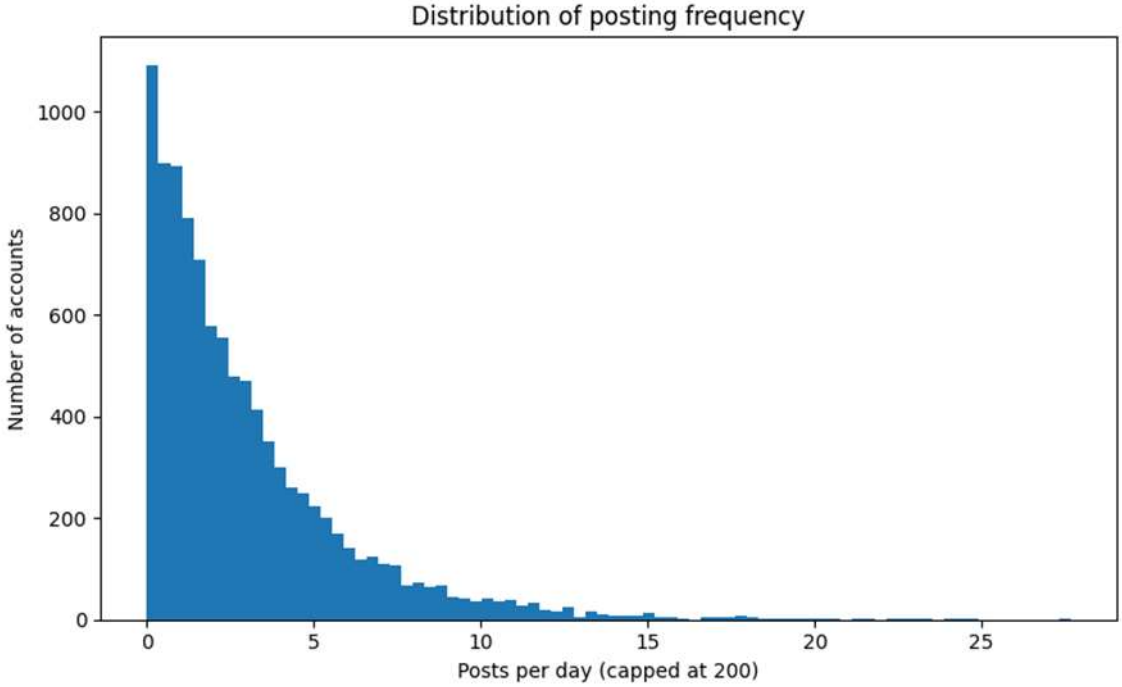
4. Results and analysis

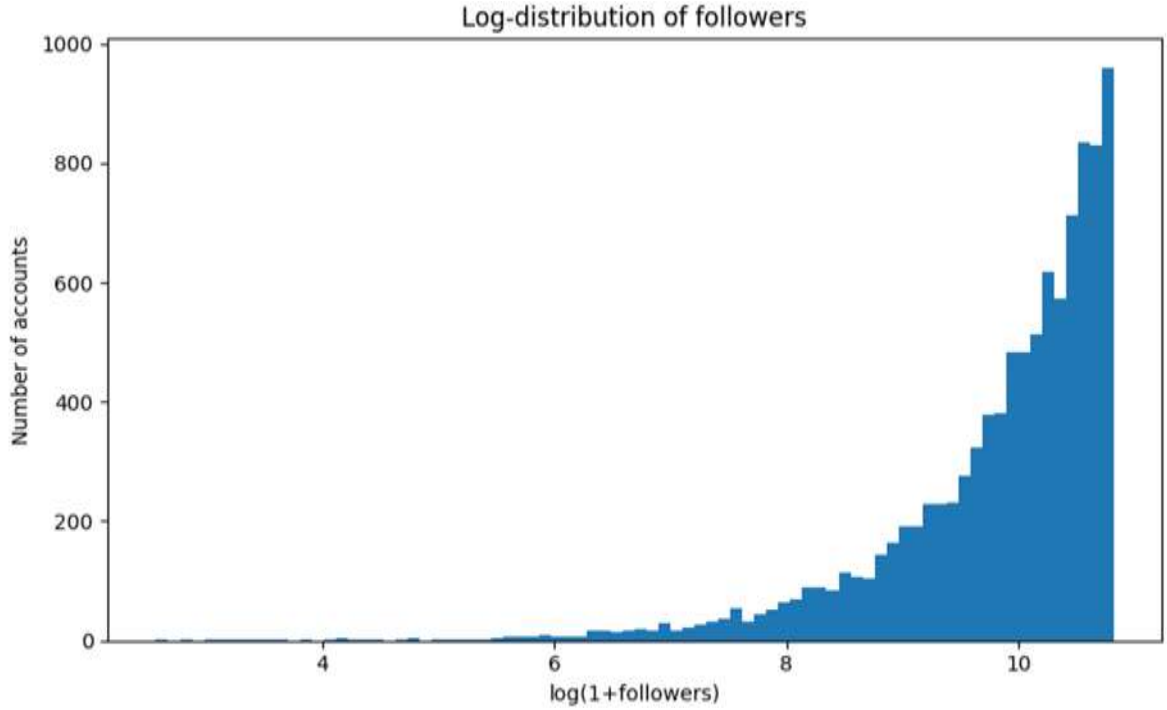
The analysis of collected social media data was conducted in three primary stages: Exploratory Data Analysis (EDA), feature engineering, and model training & testing.

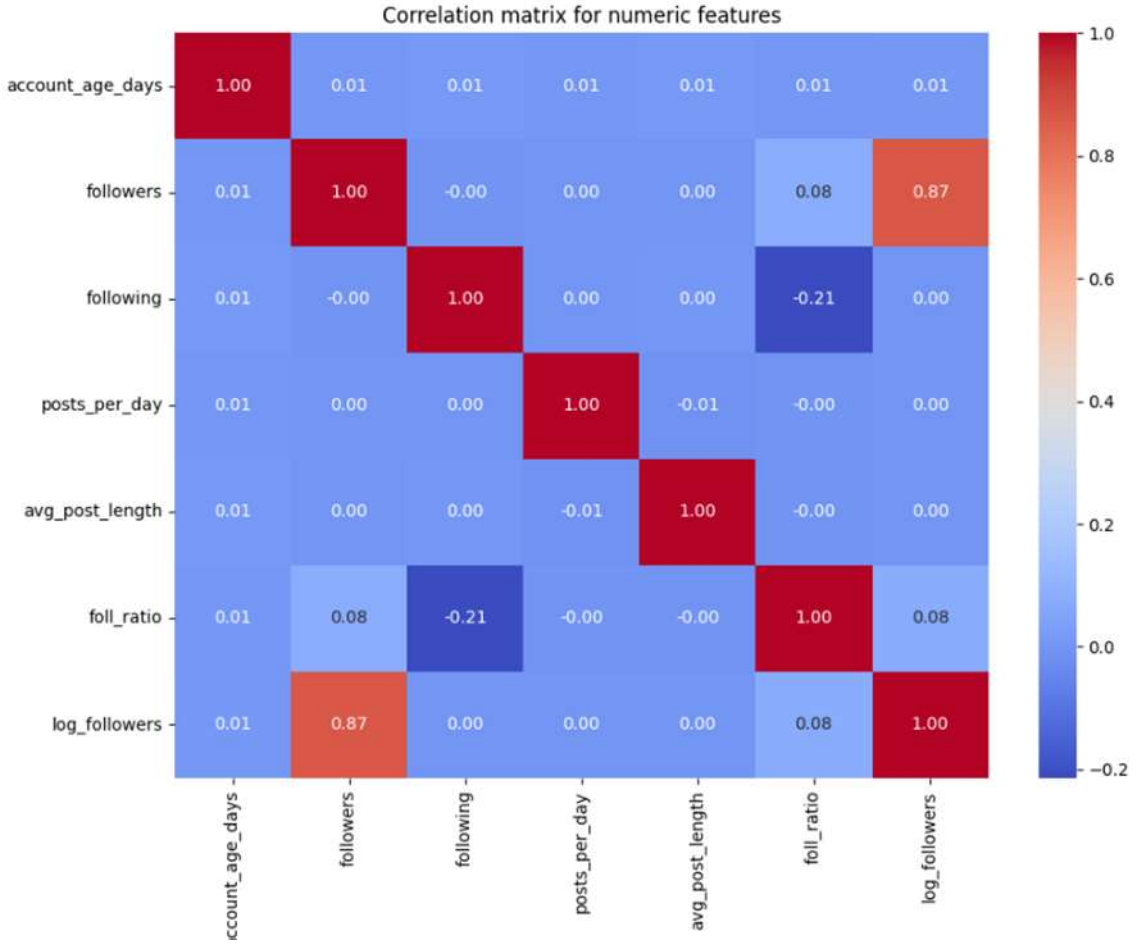
4.1 Exploratory Data Analysis (EDA)

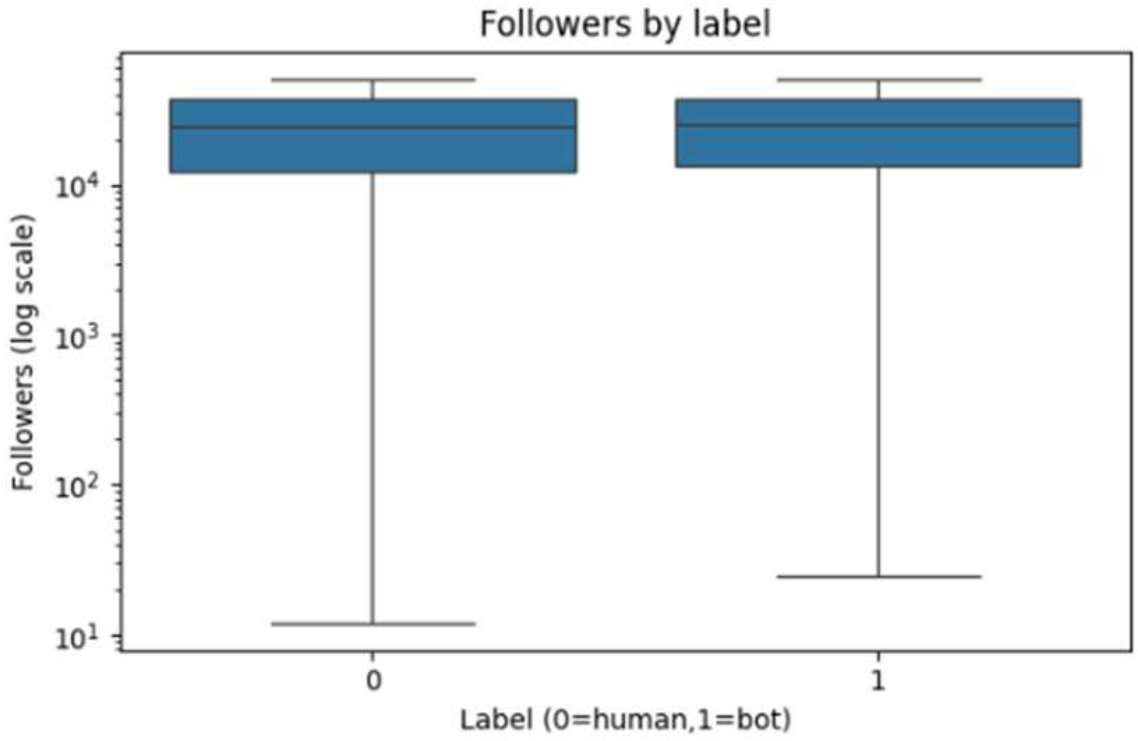
The EDA phase created an understanding regarding the distribution, correlation and anomalies existing within the dataset of about 1000,000 social media accounts of Facebook, X(Twitter) and Instagram. The results of visualisation created with Matplotlib and Seaborn showed that there were substantial differences in the account traits of real users and suspected fake ones. An example would be distributions of account ages which showed that fake accounts or automated accounts were especially underrepresented with newer and more active accounts generally having portions of missing metadata and verification badges. The analysis of networks also identified the presence of abnormalities in terms

of followers and clustering. Behavioural EDA showed the presence of unusual posting schedules, such as huge bursts of posting or a steady amount of posting.






Correlation matrix for numeric features
account_age_days
followers
following
posts_per_day
avg_post_length
foll_ratio
log_followers
1.00 0.01 0.01 0.01 0.01 0.01 0.01
0.01 1.00 -0.00 0.00 0.00 0.08 0.87
0.01 -0.00 1.00 0.00 0.00 -0.21 0.00
0.01 0.00 0.00 1.00 -0.01 -0.00 0.00
0.01 0.00 0.00 -0.01 1.00 -0.00 0.00
0.01 0.08 -0.21 -0.00 -0.00 1.00 0.08
0.01 0.87 0.00 0.00 0.00 0.08 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Followers by label
Followers (log scale)
10^4
10^3
10^2
10^1
0
1
Label (0=human,1=bot)

4.2 Feature Engineering

According to the results of the EDA, feature engineering has been used to improve the predictive accuracy of the models (Katya *et al.* 2023). Innovative composite items were also built including bot scores, behavioural and network signals, and a suspicious activity index. This includes abnormalities in the frequency of posting, engagement, and repetition of content. Correlation analysis indicated that these engineered features had a strong relationship with account authenticity, and it justifies their presence in model training. Dimensionality reduction methods were used to filter away or merge features that had a large multicollinearity, so that the models would not be biased because of conflicting inputs. Normalization and feature scaling were used when needed in order to ensure the same range across variables, especially when the algorithm is sensitive to the size of features, as in SVM and Neural Networks.

4.3 Model Training and Testing

The model training and testing were aimed at assessing the three machine learning algorithms namely the Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN) to determine fake accounts and bots on X, Facebook and Instagram data. All models trained on 80 percent of the data were cross-validated on 10-fold cross-validation in order to reduce overfitting. The grid search was used to optimize the hyperparameters to improve the accuracy and generalization. The Random Forest model had the highest balance between interpretability and performance with more than 90 percent of detection accuracy which supported H1. The Neural Network had the best recall, which validates H2 but needed more calculation and parameters optimization. SVM delivered competitive results as well as exhibited poor scalability to high-dimensional social media data. SHAP and LIME explainability analyses did show the behavioural and network properties, including the numbers of followers, the frequency of posting, and the age of the accounts. Cross-validation scores were stable in all models, which confirms the reliability of these models in the context of various data subsets. In general, the trained models developed a reproducible and effective detection model.

5. Discussion conclusion and future scope

5.1 Discussion

The results of this study prove the efficiency of machine learning methods to detect fake accounts, bots and malicious profiles on social media sites. All three models, the Random Forest, Support Vector Machine and the neural network performed highly with all of them achieving over 90 percent accuracy in the most test conditions, which supported Hypothesis H1. However, the best balance between accuracy, interpretability and computational efficiency was obtained with the Random Forest Model among them. The Neural Network model also achieved the highest recall rates as it was more capable in detecting laser or borderline malicious accounts that other models could not capture. This is in favour of Hypothesis H2, but incurred the expense of more training time and more computational load. Though the SVM worked well with smaller data sets, it was found to be limited in its scalability with high dimensional data obtained after large scale social media interactions (Ma *et al.* 2022). Such comparative observations show that deep learning is not as practical as ensemble approaches like Random Forest when analyzing large social networks. These lessons indicate the necessity of regular retraining and the renewal of features with the development of malicious actions.

5.2 Conclusion

This paper has managed to show that machine learning models are an important tool to be used in dete-

cting and countering fake accounts and bots on social media. The models created in this paper were highly reliable and practically scalable through various sources of data, sophisticated feature engineering, and significant cross-validation. The paper highlights the fact that countering online misinformation and manipulation of social media should be done with nuance and accuracy, but also with moral accountability. To ensure that people have their rights to privacy, openness in the detection algorithms, and adherence to the laws regarding data protection are instrumental to preserving societal confidence in machine learning-based moderation.

5.3 Future Scope

Although the current model is robust in accuracy and stability, there are multiple potential directions that can capture a great spot in the future. The important direction is real time implementation and combining models and in-service monitoring dashboards would potentially provide the possibility of the proactive detection of bot networks. The direction of the work in the future is the application of the cross-lingual and multimodal detection of image (Fei *et al.* 2021). The reliability of the decision-making can as well be improved by developing hybrid systems that implement machine learning and human-in-the-loop systems to combine the efficiency of the algorithms and the human judgment. Finally, ethical governance should change with technology. Fairness auditing, model transparency, and bias mitigation should be studied in the future to ensure that detection systems are no longer biased and discriminate against a particular community or communication style.

References

1. Fei, H., Yu, T., & Li, P. (2021, June). Cross-lingual cross-modal pretraining for multimodal retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3644-3650).https://aclanthology.org/2021.naacl-main.285/
2. Ghislieri, M., Cerone, G. L., Knaflitz, M., & Agostini, V. (2021). Long short-term memory (LSTM) recurrent neural network for muscle activity detection. *Journal of NeuroEngineering and Rehabilitation*, *18*(1), 153.https://link.springer.com/article/10.1186/s12984-021-00945-w
3. Hayawi, K., Saha, S., Masud, M. M., Mathew, S. S., & Kaosar, M. (2023). Social media bot detection with deep learning methods: a systematic review. *Neural Computing and Applications*, *35*(12), 8903-8918.https://link.springer.com/article/10.1007/s00521-023-08352-z
4. Ioannou, C., & Vassiliou, V. (2021). Network attack classification in IoT using support vector machines. *Journal of sensor and actuator networks*, *10*(3), 58.https://www.mdpi.com/2224-2708/10/3/58
5. Katya, E. (2023). Exploring feature engineering strategies for improving predictive models in data science. *Research Journal of Computer Systems and Engineering*, *4*(2), 201-215.https://technicaljournals.org/RJCSE/index.php/journal/article/view/88
6. LAM, G. (2025). Advanced Detection of Fake Social Media Accounts Using Machine Learning Algorithms. *INTERNATIONAL JOURNAL*, *9*(3)..https://www.ijasret.com/VolumeArticles/FullTextPDF/1468_35.Advanced_Detection_of_Fake_Social_Media_Accounts_Using_Machine_Learning_Algorithms.pdf
7. Ma, J., Xu, K., Qiao, Y., & Zhang, Z. (2022). An Integrated Model for Social Media Toxic Comments Detection: Fusion of High-Dimensional Neural Network Representations and Multiple

Traditional Machine Learning Algorithms. *Journal of Computational Methods in Engineering Applications*, 1-12.https://ojs.sgsci.org/journals/jcmea/article/view/309

8. Mohammed, S., Al-Aaraji, N., & Al-Saleh, A. (2024). Knowledge Rules-Based Decision Tree Classifier Model for Effective Fake Accounts Detection in Social Networks. *International Journal of Safety & Security Engineering*, *14*(4).https://www.researchgate.net/profile/Susan-Mohammed-2/publication/383570211_Knowledge_Rules-Based_Decision_Tree_Classifier_Model_for_Effective_Fake_Accounts_Detection_in_Social_Networks/links/66d77922fa5e11512c7341e6/Knowledge-Rules-Based-Decision-Tree-Classifier-Model-for-Effective-Fake-Accounts-Detection-in-Social-Networks.pdf
9. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 160.https://link.springer.com/article/10.1007/S42979-021-00592-X
10. Shah, A., Varshney, S., & Mehrotra, M. (2024). Detection of fake profiles on online social network platforms: performance evaluation of artificial intelligence techniques. *SN Computer Science*, *5*(5), 489.https://link.springer.com/article/10.1007/s42979-024-02839-9
11. Terumalasetti, S. (2025). Artificial intelligence-based approach to detect malicious users using deep learning and optimization techniques. *Multimedia Tools and Applications*, *84*(8), 3979-4001.https://link.springer.com/article/10.1007/s11042-024-19872-8
12. Tunç, Ü., Atalar, E., Gargı, M. S., & Aydın, Z. E. (2022). Classification of fake, bot, and real accounts on instagram using machine learning. *Politeknik Dergisi*, *27*(2), 479-488.https://dergipark.org.tr/en/pub/politeknik/issue/83819/1136226
13. Voitovych, O., Kupershtein, L., & Holovenko, V. (2022). Detection of Fake Accounts in Social Media. *Кібербезпека: освіта, наука, техніка. Т. 2,№ 18: 86-98*.http://ir.lib.vntu.edu.ua/handle/123456789/37671
14. Zhou, Y., Zheng, H., Huang, X., Hao, S., Li, D., & Zhao, J. (2022). Graph neural networks: Taxonomy, advances, and trends. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *13*(1), 1-54.https://dl.acm.org/doi/abs/10.1145/3495161