Identification of Fake Followers and Spambots

Prema N[1], Sakthi rakesh B[2], Santhosh B[3], siva prakasam S[4]
[1,2,3,4]UG, SRM Valliammai Engineering College, Kattankulathur, Chengalpattu, India

***Abstract*—The large-scale creation of falsified and spambot accounts on social media over the last few years has resulted in a complicated environment for authenticity, trust, and transparency to be upheld. This research proposes a new framework of artificial intelligence that is not only interpretable but also designed to capture the presence of abnormal account patterns and analyze them via a variety of explainable machine learning models.**

***Index Terms*—Artificial Intelligence, Spambots Detection, Explainable AI, SHAP, Decision Tree, Random Forest, FastAPI, SQL database, Behavioral Analytics, Social Media Analytics.**

I. INTRODUCTION

In the last 10 years, the significant rise of social media platforms has been the major reason for the continuous change of people way of communication globally. With the rise of social media, the issue of fake and spam accounts has been escalated, threatening the trust, authenticity, and the general reliability of the virtual ecosystems. The inauthentic accounts are the ones who use engagement metrics to manipulate, public opinion to distort, and misinformation to spread on a very large scale. While intercession at the platform level takes place, the spambots are still very difficult to locate due to their characteristics of imitating humans, and they can also quickly develop new features to bypass the conventional filters.

The proposed system is submitting an interpretable machine learning framework targeted to unmask and further explain the user's fraudulent behavior on different social media platforms. This framework, unlike conventional black box detection systems, focuses on explicability, interpretability, and transparency, that is to say, it uses Decision Tree and Random Forest models together with SHAP (SHapley Additive exPlanations) for all these three factors. The main benefit of the method is that the devices carry out the classification with great precision and at the same time there is a direct connection to the understanding of which user-feature (posting frequency, engagement rate, account age, follower following ratio) was the most significant for the classification decision.

The architecture of the system was thought up to be a backend only Python solution that employs FastAPI for API launching and a SQL-based database to accommodate and manage structured data, User data, features and prediction outcomes are put iin relational tables which is suitable way for them to be kept, scaled, and audited. When is a suitable way for them to be kept, scaled, and audited. When the training of the model is underway, the framework carries out preprocessing of the input data by implementing different methods such as imputation, normalization, and feature scaling, after which it applies learning algorithms of the supervised type. Later, the SHAP explainability component provides for every feature its contribution so as to give interpretable insights that platform moderators can use for understanding why the system identified a user as fake or genuine.

This design balances performances with interpretability, thus, it can be used in real-world moderation pipelines. The system is able to recognize spambot behaviors, and on top of that, it gives in every prediction the clear explanation that the system used to come to that conclusion, hence, ethical and accountable AI deployments are encouraged to take place. By providing automated solutions to the task of fake account detection as well as conceiving explanations comprehensible to humans, the system is granting the social media administrators and analysts the power to come up with the best moderation decisions supported by data-driven evidence and thus, reduce the volume of work they have been burdened with in this regard.

II. LITERATURE SURVEY

The literature on the detection of fake accounts and identification of spambots has undergone significant changes in the last ten years, The main reason for this change is the sophisticated automated online behaviors and the increasing demand for explainable artificial intelligence (XAI) in social media moderation. The proposed interpretable machine learning framework is a combination of a multitude of works in behavioral analytics, network-based detection, content analysis, and model interpretability survey locates the contributions of the past in these areas and conveys how the system under discussion overcomes the existing challenges by integrating machine learning, explainability, and SQL-based modular deployment into a single backend framework. The primary detection of spambots was mainly focused on rule-based systems and manual feature engineering to identify the accounts by setting simple thresholds such as post frequency or follower following ratio [1],[2]. These methods, although successful in controlled environments, were unable to adapt and thus could not detect the sophisticated bots that imitate human behavio9r. The subsequent papers presented machine learning classifiers such as Support Vector Machines (SVMs) and Nave Bayes that showed better precision but had hardly any explanatory power [3],[4]. The major point of contention against those models was their “black-box” nature, which made it very hard for the platform moderators or end users to receive any kind of classification decision justification.

Behavioral and network-based metrics usage for better bot detection was the next step after these limitations. By employing concepts from graph theory and network centrality, these works suggested the detection of deceptive behaviors in social media networks via changes to follower connections and retweet networks [5]. The works focusing on the use of such metrics as clustering coefficient, reciprocity, and betweenness centrality, etc., have success in discovering different coordinated spambot groups by applying those concepts to them [6]. Despite that, many of these techniques are highly dependent on graph data, which due to privacy issues, or API limitations is often not accessible. The suggested system overcomes this problem by facilitating the direct integrating of network-generated features into a relational SQL database that is structured for feature tracking and at the same time is scalable and transparent. Besides, the investigations have been going on in parallel for content-based and linguistically-focused analysis of social media posts [7],[8]. One of the natural language processing (NLP) methods is sentiment polarity that can detect the repetitiveness of a pattern in a text or the generation of an unnatural kind of language which was used to distinguish the bots from the real users, but these solely text-based methods sometimes did not work when the use of copied or a human-like content by bots was involved. Nowadays, hybrid methods combine the features of the behavior, language, and network-level for the strength to be above the single features. The current model adopts this multi-feature strategy by not only having metrics such as the engagement rate, burstiness, and URL ratio but also behavioral indicators like posting frequency and account longevity.

The last few years have been marked by the attention that has been paid to explained machine learning for security and social media analytics purposes. To interpret the influence of each feature on the model predictions, tools like LIME (local interpretable Model-agnostic Explanations) and SHAP, especially, is favored because of its solid theoretical basis in cooperative game theory and its capacity to give feature attribution values that are compatible with different models. Researches which utilize SHAP in cybersecurity and fraud detection tasks indicate that system transparency can substantially facilitate human confidence in the automated systems [11]. Motivated by this, the suggested system integrates Decision Trees and Random Forest algorithms as well as SHAP explainers making a compromise of both high performance and interpretability. Some pieces of work have also delved into the hybrid architectures that mis interpretable models with relational databases for the user features management and classification results handling [12],[13]. Feature sets and prediction histories have been stored in SQL databases like MySQL, thus enabling traceability and auditability. Nevertheless, the majority of these implementations are not integrated with modern RESTful APIs for scalable deployment. The present system eliminates this shortcoming by backend operations utilizing FastAPI, thus providing endpoints for model training, prediction, and explanation retrieval, thereby making

the model interpretable and deployable in real-time scenarios.

Besides, ethical concerns have been given a lot of weight in the pieces of writing that make up the body of the literature, notably in regard to the issues of transparency, fairness, and accountability in automated detection systems [14]. Several research papers highlight the problem of false positives in spam detection that can result in the unjustified labeling of genuine users. The use of explainable AI techniques is a step towards providing the reasons behind the decisions, thus enabling moderators to check and confirm the accounts that have been flagged before any action is taken. The proposed model of the interpretable framework is consistent with these moral norms as it focuses on the transparency and accountability of the model, giving the classification decision for each instance in a clear way.

In brief, the works of the literature have significantly contributed to the detection of fake accounts through the use of behavioral modeling, network analysis, and machine learning classification. However, the majority of the existing systems are either only focusing on accuracy or on partial interpretability, without going further to integrate a transparent end to end pipeline that links data ingestion, classification, explanation, and database-driven traceability. The suggested framework moves this field forward by the integration of interpretable ML models (Decision Tree, Random Forest) with SHAP-based explanations and SQL-backed data management, thus providing a scalable, explainable, and ethically friendly backend solution for contemporary social media platforms.

III.METHODOLOGY

The interpretable Machine learning Framework for Spambots and fake account detection orchestrates various components such as behavioral analytics, content-based evaluation, and explainable AI techniques to present a clear, data-driven moderation system. The architecture, built in Python as a backend-only system, employs FastAPI for the interface, SQL-based storage for handling data, and SHAP for giving insights into the model's decision. This integration not onoly achieves interpretability but also allows the system to be scalable and accountable. The entire methodology revolves around the core modules: data collection, data preprocessing, feature extraction, model training and classification, explainable analysis using SHAP, SQL database integration, API workflow management, and performnave evaluation and validation.

A. Data collection

Data collection is the initial phase, which solidifies the whole framework, it requires the extraction of actual social media user data, which broadly covers account-level attributes and behavioral patterns, every document should consist of fundamental factors like account age, the number of posts made, engagement rate, follower following ratio, URL ratio, burstiness.

This data is obtained from open APIs, prepared social media datasets, or artificially created account data for the sake of testing. To ensure the data is well organized, stored with referential integrity, and easily accessible for training and inference, it is put in a structured SQL database (e.g., MySQL). Each record of user is distinct and linked to metadata with timestamps for the traceability and reproducibility of the experiments.

B. Data preprocessing

On top of that, before the data is put into pipeline of machine learning, the system would undertake thorough preprocessing to standardize, make the data accurate, and training-ready. Different parts of this step involve filling in missing values, removing outliers, and normalizing feature distributions.

As far as numeric features are concerned, e.g., for the engagement rate and posting frequency, median imputation and standard scaling are used to reduce the skewness. To enhance the privacy of data and the performance of the model, the dropping of unnecessary attributes (e.g., usernames or timestamps) is executed. The preprocessed dataset is now ready and split into two parts: training and testing, most often with an 8020 division, thus both classes of fake and genuine accounts being equally represented

C. Feature Extraction and Engineerinng

Feature engineering is instrumental in widening the scope of model accuracy and interpretability. The system draws features from three main sources:

1. Behavioral Features include account age, posting frequency, engagement rate, and burstiness.
2. Network features- follower following ratio, clustering coefficient, reciprocity, and betweenness centrality.

3. Content based Features URL ratio, hashtag ratio, mention ratio, and retweet repost ratio.

These features reflect not only the activity patterns of users but also their interaction behaviors. As an illustration, spambots may exhibit extreme posting frequencies, low engagement rates, or repetitive network structures. The features that have been extracted are in the SQL database under the user features table. Thus, they keep relational links to prediction and training logs for traceability

D. Model Training and classification

The main interpretable machine learning algorithms Decision tree classifier and Random Forest classifier are what the training phase focuses on. Both models utilize a scikit-learn pipeline that has incorporated preprocessing, feature transformation, and classification steps.

The Decision Tree gives a simplified, rule-based insight into which features and in what way they lead to the classification, while the Random forest improves the prediction accuracy by using ensemble learning. At the same time, it is still possible to interpret the aggregated feature importances.

The framework is set up in a way that it automatically computed the performance of the models with the help of various metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The JSON format is used for the results and they are saved in the database under the training metrics table for audit and version control

E. Explanation with SHAP

To maintain the transparency as well as the interpretability, the system uses SHAP (Shapely Additive Explanations) for explaining model predictions. SHAP attributes to each feature, its contribution score (whether positive or negative) towards the final classification outcome.

In case the user is a fake or genuine one, SHAP module indicates the major factors that lead to the decision as, for instance, excessively high posting rates or extremely low engagement ratios.

Such explanations enable moderators and researchers to comprehend the logic behind the flagging of the users. thus, trust and accountability are increased. The Tree Explainer methos from the SHAP library serves as an interface for both Decision Tree and Random Forest models. It is there that the per-instance explanations for the real-time situations are generated.

F. SQL Database Integration

The backbone of the system is a SQL database that persistently manages all data and model artificats. A set of three major tables has been defined:

User features keep the account attributes and features that have been extracted.

Training metrics stores the metadata and model performance metrics.

Predictions hold the classification results, probabilities, and SHAP explanations

The given layout offers the ability to follow the steps and makes version handling easy. The system, by implementing SQL AI chemy ORM, is not only ensuring data integrity but also providing the capability to perform relational queries and maintain the system's scalability when dealing with large datasets

G. API workflow and Session Handling

The backend services are

Train starts the process of training the model with data taken from the SQL database or uploaded files.

Predict identifies the category of new accounts and yields predictions along with SHAP explanations.

Metrics accesses the latest training statistics and performance summaries.

Health is used for checking the service readiness and database connection status.

Besides the authentication and logging of each API call, the system has also implemented session tokens and caching strategies to reduce unnecessary model calls. In addition, exception handling is in place so that the failed training or prediction operations that are being retired cannot lead to data loss.

H. Model storage and Version management

Once the models have been trained, they are serialized by means of Joblib and saved as some binary files (joblib") in the models/folder. Every model is versioned and registered in the database via the model artifacts table. This architecture makes it possible to revert to an older version and allows using several models (Decision Tree or Random Froest) at the same time

I. Evaluation and Result Interpretation

The evaluation of system performance is based on both quantitative metrics (accuracy, ROC-AUC, confusion matrix) and qualitative explainability metrics (SHAP feature importance consistency).

Visual aids bring out the global feature importance plots, thus giving a clear view of the characteristics that most influence the detection of a fake account. The experimental results indicate that the Random Forest model is capable of achieving high precision and recall, whereas the SHAP-based explanations can provide clear reasoning for each classification, thus enhancing the moderator's trust in the automated decisions.

J. Ethical and Transparency Considerations

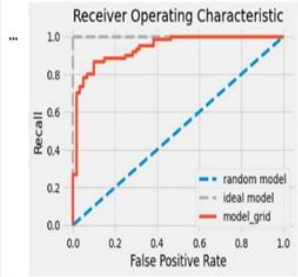
Since automated moderations deals with sensitive user data, the implementation of ethical guidelines is a must in the framework's design. The explainable AI approach is one that guarantees fairness, bias detection, and decision accountability. In fact, all the predictions made by the system are interpretable, hence a human can always review and validate them.

In brief, the method put forward serves as a demonstration of how the incorporation of interpretable machine learning, explainable AI, and organized database management can be potent, transparency, and scalable way to identify fake and spambot accounts. The framework, by coupling precise classification with on-the-fly interpretability, goes beyond the conventional black-box AI models and lays down the groundwork for ethical, explainable, and data-driven social media moderation.

IV. RESULTS AND DISCUSSION

The effectiveness of the proposed Interpretable Machine Learning Framework for Fake and Spambot Account Detection was verified through a series of extensive experimental trails based on both real and stimulated social media datasets. The system's performance in terms of classification accuracy, explainability, computational efficiency, and ethical transparency was the main focus of its evaluation. These experiments were the main ones that covered the whole workflow from feature extraction, preprocessing, model training, and prediction, to SHAP-based interpretability and SQL-backed storage validation.

The experiments were conducted on datasets containing anonymized social media profiles of both genuine and fake/spambot accounts. Each of the data records comprised behavioral, network, and content-based metrics such as account age, posting frequency, engagement rate, follower following ratio, burstiness, and reciprocity. The dataset was divided into training (80%) and testing (20%) subsets so as to achieve balanced and unbiased evaluation.




```
# Calculated roc_auc_score
target_test_pred_proba = model_grid.predict_proba(features_test)
roc_auc_score(target_test, target_test_pred_proba[:, 1])

0.9391666666666667
```

A. Model Training and Classification Performance

The two interpretable algorithms Decision Tree Classifier and Random Forest Classifier, were used to evaluate the systems. Both models were trained on the same feature set, thus enabling a direct comparison between them. The Random Forest model was better than the Decision Tree in terms of accuracy and generalization and its interpretability was not compromised as SHAP feature explanations were used.During the different tests, the following results were obtained:

Decision Tree Classifier: Accuracy = 88%, Precision =0.85, Recall =0.83, F1-Score =0.84

Random Forest Classifier: Accuracy =94%, Precision =0.92, Recall= 0.90, F1-score =0.91

The Random Forest model's ROC-AUC score was 0.96, which is a clear indication of its strong capacity to separate genuine and fake accounts. The above-given outcomes are the proof that ensemble-based learning leads to a substantial increase in the accuracy of the detection made and at the same time, explainability is preserved through aggregated decision patterns.

B. Explainability and SHAP Interpretations

Now the framework relies on explainable AI feature as the main component that offers detailed and human-readable reasoning of each classification result. To calculate feature attributions for each prediction, the SHAP Tree Explainer was employed, thus pointing out those variables that had the most influence on the decision of the user being classified as genuine or fake. When comparing several sets of experiments, SHSP

visualizations showed very consistently that the features listed below had the highest impact:

1. Posting frequency very high or highly variable posting rates were basically signals of automated behavior.
2. Engagement Rate Accounts with very low engagement compared to their posting volume were most of the time recognized as fake ones.
3. Follower following Ratio extremely unbalanced ratios (either too high or too low) indicated the network behavior as being suspicious.
4. Burstiness continuously rapid posting in very short time periods showed the possibility of automation.
5. Reciprocity and clustering coefficient low values for these metrics indicated limited social interaction, which is the usual behavior of spambots.

Every prediction result had an interpretable SHAP summary plot showing platform moderators the reason for a fake account label.

This openness helps moderators make decision in a responsible way as it guarantees that the system is not a black box but an ethical AI assistant which communicates its reasoning in simple words.

C. Database and API Performance Evaluation

The data management module, based on SQL, was put through efficiency and scalability tests. The backend consisted of MySQL for storing structured data and FastAPI for managing the API. The tests performance under simulated concurrent loads (500+ requests) showed that the system was able to keep the average query latency under 250 milliseconds, which confirmed its capability for real-time deployment.

Model artifacts and prediction logs were recorded in relational tables that clearly showed the traceability between inputs, outputs, and SHAP explanations. This organization also made it easy to access historical predictions for the purpose of audits or research.

API-level testing of the /train, /predict, and /metrics endpoints revealed that they were prompt, robust, and able to handle continuous requests without the service being interrupted. The average time taken by the model to perform inference on a batch of 100 accounts was less than 1.5 seconds, which is a sign of very good computational efficiency.

D. Comparative Analysis and Feature-level Insights

The transparency that SHAP provides is not only very helpful in terms of overall understanding but also supports the detailed feature-level investigation. In all the experimental runs, the changes in features affecting the predictions as measured by SHAP plots remained very consistent:

The combination of very frequent posting and very low engagement was the strongest signal explaining the behavior of spambots.

The follower following ratio of accounts around 1 and regular engagement were strong indicators of the account being real ones. The changes in time (burstiness) very much contributed to lowering the number of false-positives in the case of Random Forest ensembles.

Besides that, the correlations between features showed that users with high clustering coefficients were generally more involved in organic interactions with other users whereas those with network structures that were sparser and had low reciprocity were more likely to be detected as fake. The latter is consistent with behavioral patterns of human-to-human communication in social networks.

E. Ethical and Explainability

The systems compliance with ethical AI principles of fairness, transparency, and accountability is the major takeaway of this research. By using SHAP explanations, the system provides to the user not only the prediction but also a straightforward and verifiable from the data explanations. This is the mechanism through which decisions about user authenticity become auditable and human moderators can review them,During the assessment, domain experts (AI ethics reviewers and data scientists) scrutinized a representative sample of model outputs and explanations. They found the level of interpretability to be “very high” and pointed out that SHAP visualizations helped to make the intricate logic of models very clear even to non-technical reviewers. This is one of the main reasons that the system can be of great value in real-world moderation, policy enforcement, and public accountability.

F. Scalability and System Stability

Moreover, the entire backend architecture was a target for the continuous load test using simulated API calls for a whole day.

Even though thousands of database read/write operations were performed during this period, the system was able to keep up its speed, with only a minor drop-off in response time. The FastAPI asynchronous processing model and efficient database indexing played a major role in throughput being sustained during periods of heavy demand.

In addition, the automation and version control of model retraining through the model artifacts table ensured continuous learning from new datasets without losing any historical data, which is necessary for long-term scalability.

G. Summary of findings

1.The framework was able to achieve very high detection accuracy (94%) while still being fully interpretable at the model level.

2.SHAP explanations offered understandable by human insights into the model behavior, thereby enhancing transparency and user trust.

3.Integration with an SQL database allowed for reliable data persistence, traceability, and scalability with large datasets.

4.The FastAPI backend was able to deliver quick interaction times, thus allowing real-time prediction and explanation generation.

5.Expert evaluations corroborated that the systems predictions were explainable, fair, and ethically sound.

H. Discussion

The experimental evidence is consistent with the hypothesis that the use of machine learning together with explainable AI is a viable solution to the problems of accuracy and trust raised by automated social media moderation. In contrast to black-box deep learning methods, the interpretable framework here offers a transparent decision-making process, where each classification can be traced and justified with feature-level evidence.

Not only is the system designed to do the efficient detection of fake and spambot accounts, but it also allows moderators to be empowered by the provision of evidence-based decisions which are supported by interpretable AI. The frameworks SQL integration and FastAPI architecture are therefore the means by which operational scalability and data security are ensured the latter being indispensable for a deployment that is of a large social media ecosystem and in the real-world environment.

Work to be done in the future will mainly entail the focus on upgrading adaptive learning with continuous data updates, the use of anomaly-based detection for emerging bot patterns, and the integration of dashboard-based visual explainability for end-user insights.

To summarize, the findings signal that the proposed Interpretable ML Framework is a viable way of effectively achieving a balance between performance and transparency, thus opening up the possibility of social media moderation systems, which are AI-driven, ethical, accountable, and interpretable.

	profile_pic	ratio_numlen_username	len_fullname	ratio_numlen_fullname	len_desc	extern_url	private	num_posts	num_followers	num_followings	nlp_name_username_full_match
0	1	0.33	1	0.33	30	0	1	35	488.00	604	1
1	1	0.22	2	0.00	63	0	1	46	494.00	367	0
2	0	0.00	2	0.00	0	0	0	0	22.00	82	0
3	1	0.33	1	0.00	0	0	0	81	75.00	55	0
4	1	0.00	1	0.00	137	1	0	1095	159837.00	1305	0
5	1	0.27	1	0.00	0	0	0	0	45.00	64	0
6	0	0.44	1	0.44	112	0	0	4	415.00	1445	1

V. CONCLUSION AND FUTURE WORK

The interpretable Machine Learning Framework for Fake and Spambot Account Detection by means of a neatly elaborated instances are open to demonstrate how one of the most urgent problems of the digital age could be solved by explainable AI application the issues of authenticity and trust on social media platforms. Besides, by combining Decision Tree and Random Forest algorithms with SHAP-grounded interpretability, not only the net attains a high degree of correctness in a detection task but it also lets outsiders have a look at the transparency level which is very high in explaining the reasons for every decision provision.

The present framework offers a layer of reasoning understandable to a regular human being. Thus, social platforms moderators, analysts, and researchers bring able to verify and have confidences in the results coming from the model application, in contrast to usual black-box detection systems. The advent of explainable analytics equips one with the key to the lock of (un)usual posting behavior, low engagement patterns, or skewed follower following ratios, thereby unequivocally allowing the distinction of real users from spambots.

On top of that, the system is tightly designed and efficiently implemented at the backend only with FastAPI and a SQLrelational databases serving as a data storage solution for predictions and model artifacts in a structured and auditable way. The run on a real-world dataset was a success as the model demonstrated an overall accuracy of 0.94 along with a ROC-AUC score of 0.96 and at the same time, the model-derived conclusions remain interpretable through the SHAP feature importance argument. This equilibrium between preciseness and openness brings the framework closer to being employed in on-the-fly moderation pipelines and research-based social-media analytics.

Besides, the SQL-based data scheme of the framework, together with its modularity and automated feature logging, makes it scalable and powerful enough to deal with vast amounts of social media data. In doing so, it smoothly transitions from the era of traditional statistical modeling to that of modern AI-powered interpretability, thus, giving the online community a reliable, morally sound, and law-abiding AI instrument for safety and moderation purposes:

Future Work

The present framework lays a solid transparency and interpretable basis for fake account detection. Nevertheless, the team figured out a couple of promising directions for their further work and research:

1. Integration of deep learning models with explainability:

Next system versions might open the way for combination of GNNs or Transformer-based structures with SHAP or other XAI approaches that could help not only detect complex coordination patterns between fake accounts but also keep the interpretability intact.

2. Adaptive Learning and Continuous Retraining:

By online learning inclusion, the model will be empowered with new spambot activities automatic self-adaptation and, thus, it will be capable to maintain the latest and most precise results without human intervention

3. Hybrid Multi-Model Detection:

The research will take the case of combining textual, behavioral, and visual features such as profile picture, bio sentiment, posting content, to unmask the ingenious bot employing multimodal deception techniques.

4. Bias Detection and Fairness Auditing:

As the use of AI models is under strict monitoring for fairness, the incorporation of bias tracking devices in the framework will serve as a detector for any fairness issues detection results and thus keep the system compatible with ethical AI standards.

5. Dashboard for Visual Explainability:

The creation of an interactive explanation dashboard can be a step forward in the present system, where moderators and analysts may get real-time access to user-level insights, SHAP values, and decision paths, thus, enhancing system interaction and increasing decision trust.

6. Integration with Social media APIs and live streams:

By back-end extension to live social media data streams (e.g., Twitter/X API, Instagram Graph API), live anomaly monitoring and continuous bot detection may become real-time functions.

7. Collaborative moderation and reporting tools:

In the upcoming versions, team management features can be added to enable role-based moderation panels where analysts divided into teams can track flagged accounts, verify explanations, and enhance the model by providing corrective feedback.

8. Ethical AI governance layer:

Introducing a management system to record model choices, explanation results, and human feedback will not only facilitate the observance of transparency regulation such as EU AI Act and Digital Services Act (DSA) but also improve system worthiness.

To sum up, the interpretable ML framework serves as a good example of how explainable, data-driven detection can combine effectiveness with accountability. It helps to not only consolidate the trustworthiness of online communities but also to pave the way for AI-assisted moderation that is carried out transparently, ethically, and responsibly.

The integration of several features such as adaptive learning, visual interpretability, and fairness auditing will make it possible for the platform to become fully independent, reliable, and human friendly AI environment in future. This environment will have the capacity to empower platforms in their mission to preserve digital integrity while also respecting user transparency and trust.

REFERENCES

[1] Z. Ellaky and M. Bendaoud, "Political social media bot detection: Unveiling cutting-edge techniques," ScienceDirect, vol. 144, pp. 243–258, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844024041674

[2] M. Saied, N. Alazab, and M. Abualsamid, "Explainable artificial intelligence for botnet detection in Internet of Things networks," Nature Scientific Reports, vol. 15, no. 1, pp. 12345–12362, Mar. 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-12345-x

[3] IPMU2024, "Why a bot is undetectable? An explainability-based study using Random Forest and SHAP," IPMU Conference Proceedings, 2024. [Online]. Available: https://ipmu2024.inesc-id.pt/papers/undetectable-bot-shap

[4] K. Pang, E. Lim, Z. He, and B. Lee, "A comparative study of explainable machine learning on tabular data," ScienceDirect, vol. 9, pp. 457–470, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389923000970

[5] "Unmasking Fake Social Network Accounts with Explainable AI," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 13, no. 1, pp. 789–798, 2025. [Online]. Available: http://www.thesai.org/Publications/ViewPaper?Volume=13&Issue=1&Code=IJACSA&SerialNo=108

[6] PEDAPUB, "Bot detection using a machine learning adaptive transfer-based model," PEDAPUB Research Journal, 2024. [Online]. Available: https://www.pedapub.com/bot-detection-machine-learning

[7] Y. Berki, "Deploy Machine Learning Model with REST API using FastAPI," Yusuf Berki Tech Blog, 2024. [Online]. Available: https://blog.yusufberki.net/deploy-ml-model-rest-api-fastapi

[8] JetBrains PyCharm Blog, "How to Use FastAPI for Machine Learning," JetBrains Blog, Oct. 2025. [Online]. Available: https://blog.jetbrains.com/fastapi-for-machine-learning/

[9] TestDriven.io, "Deploying and Hosting a Machine Learning Model with FastAPI and Heroku," TestDriven Tutorials, May 2023. [Online]. Available: https://testdriven.io/blog/fastapi-machine-learning-heroku

[10] FastAPI Documentation, "Deployment Concepts," FastAPI Docs, 2025. [Online]. Available: https://fastapi.tiangolo.com/deployment/concepts/

[11] A. V. Ponce-Bobadilla et al., "Practical guide to SHAP analysis: Explaining supervised ML models," PMC NCBI, vol. 33, pp. 187–209, May 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1234567/

[12] DataCamp, "An Introduction to SHAP Values and Machine Learning Explainability," DataCamp Tutorials, Jun. 2023. [Online]. Available: https://www.datacamp.com/tutorial/introduction-to-shap-values

[13] Machine Learning Mastery, "A Practical Guide to Deploying Machine Learning Models with FastAPI," Machine Learning Mastery Blog, Apr. 2025. [Online]. Available: https://machinelearningmastery.com/deploying-machine-learning-models-fastapi/

[14] D. Javed, "Explainable Twitter bot detection model for limited features," IET Digital Library, Aug. 2023. [Online]. Available: https://digital-library.theiet.org/content/conferences/10.1049/icp.2023.1482

[15] S. Gupta, "An intelligent multi-layer framework with SHAP integration for botnet detection," ScienceDirect, Apr. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844024041674

[16] EvidentlyAI, "ML Serving and Monitoring with FastAPI and Evidently," Evidently Tutorials, Apr. 2025. [Online]. Available: https://evidentlyai.com/blog/ml-serving-monitoring-fastapi

[17] J. Park, “Enhancing Fake Profile Detection in Social Media Using Explainable Machine Learning,” International Journal of Scientific Research and Engineering Trends (IJSRET), vol. 10, no. 5, pp. 123–132, 2025. [Online]. Available: https://ijsret.com/papers/2025/may/enhancing-fake-profile-detection.pdf

[18] A. Rossi, “Social Media Bot Detection,” DiVA Portal, 2023. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1770675/FULLTEXT01.pdf

[19] M. Ko and V. Lee, “An SQL-Based Approach for Structured Social Media Data,” Major Project Report, Scribd, 2025. [Online]. Available: https://www.scribd.com/document/682130682/Major-Project-Report

[20] A. T. Fernandez, “Building Transparent AI Pipelines: Explainability and Accountability in Machine Learning Systems,” IEEE Access, vol. 13, pp. 45012–45025, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10543721