An Intelligent System For Identifying Fake Social Media Accounts

Mrs. D.Himabindu[1], Sharoni Hashika Pentakota[2], Prasadi Siddhardha[3], Vudi Sudhakar[4], Peteti Vijay Kumar[5]
[1,2,3,4,5]Department of Computer Science and Engineering (Data Science)
Raghu Institute of Technology, Visakhapatnam, India

Abstract

The rise of spambots and fake followers on social media platforms poses significant challenges to the authenticity of online interactions. Fake accounts manipulate engagement metrics, spread false information, and distort online conversations, undermining the credibility of these platforms. This project aims to address this issue by utilizing machine learning algorithms combined with interpretable AI techniques to effectively identify and classify spambots and fake followers. The dataset used for model training includes various features such as user interactions, post content, engagement metrics, and other relevant metadata. For classification, the project employs several machine learning algorithms, including K-Nearest Neighbors (KNN), Bagging Classifier, Stacking Classifier, and CatBoost Classifier. To enhance interpretability, Partial Dependence Plots (PDP) are applied, allowing a clear understanding of how individual features influence model predictions. The evaluation of model performance is conducted using standard metrics such as accuracy, precision, recall, and F1-score, ensuring the models' robustness in distinguishing legitimate users from fake accounts. The project's primary objective is to develop an automated solution that can accurately identify fake followers and spambots, thereby improving platform transparency, credibility, and user trust. Ultimately, the research will contribute to the development of more efficient methods for detecting fake activities, benefiting social media platforms and their users.

Keywords: Spambots, Fake Followers, Machine Learning, Classification, Partial Dependence Plots, CatBoost, BaggingClassifier, K-Nearest Neighbors, StackingClassifier, Web Interface

1. Introduction

The rapid growth of social media platforms has brought about a surge in fake accounts, often referred to as spambots and fake followers, which disrupt genuine user engagement and inflate interaction metrics. These fake accounts are used to manipulate platform algorithms, spread misinformation, and even distort social trends. To address these concerns, machine learning-based detection systems have gained significant attention as they offer automated and scalable solutions for identifying and classifying fake accounts based on user interactions, post content, and engagement metrics. However, these systems face challenges such as ensuring high detection accuracy, interpretability of the models, and handling large datasets with diverse behaviors and patterns of fake accounts.

The novelty of this work lies in the use of multiple machine learning models, including K-Nearest Neighbours (KNN), Bagging Classifier, stacking Classifier and CatBoostClassifier, combined with interpretable AI techniques such as Partial Dependence Plots (PDP) to enhance model transparency. The use of these advanced models allows for the accurate classification of fake accounts while providing insights into the influence of various features on the predictions. The ability to interpret the model's decision-making process is crucial for developing trust in automated systems for fake account detection. By leveraging the power of ensemble learning and interpretability, the system can address challenges such as imbalanced data and feature complexity, ensuring high accuracy in detecting fake followers while maintaining transparency for platform administrators.

This approach goes beyond traditional methods by offering not only classification but also interpretability, which allows platform administrators to make informed decisions based on model outputs. With the integration of these models, the system can be adapted to scale for large social media networks, providing an efficient solution for improving platform credibility and user trust. Moreover, real-time processing capabilities enable the system to flag fake followers and spambots promptly, preventing further disruptions and fostering a more authentic online experience.

The effectiveness of these machine learning models is contingent upon the quality of training data, feature optimization, and the ability to manage the ever-changing nature of online interactions . This research underscores the importance of continually refining model performance and integrating advanced data augmentation techniques to improve the robustness of fake account detection systems. Ensuring the accuracy and scalability of these systems is essential for combating fake followers and spambots on social networks.

2. Literature survey

The rise of spambots and fake followers on social media platforms poses significant challenges to the authenticity of online interactions. Fake accounts manipulate engagement metrics, spread false information, and distort online conversations, undermining the credibility of these platforms. Among other research works, the focus on enhancing detection accuracy has been pursued by applying sophisticated techniques in user data and integrating machine learning models. The different models rely on varying social media data, which comprise user interactions, post content, engagement metrics and other relevant metadata, in identifying and classifying spambots and fake followers.

The researchers of a research created a machine learning framework, where K-Nearest Neighbors (KNN), Bagging Classifier, Stacking Classifier and CatBoost Classifier are applied, to identify spambots and fake followers using multiple features. The model they made had high accuracy in classification, indicating the capability of machine learning in detection problems of social media accounts. The hyperparameter tuning and class imbalance handling methods were applied in the model, which is a widespread issue in social media datasets as presented in [1].

Machine learning was also applied by an alternative method to classify fake accounts by analyzing user interactions, post content and engagement metrics. The system created an ensemble model which allowed high performance on various types of features and strong validation accuracy. The paper used proper optimization techniques to train the model with the training stability being enhanced with the help of regularization as presented in [2].

The entire analysis of machine learning methods in the detection of spambots showed that ensemble models with fine-tuning gave effective results regarding the classification of fake followers. Their results showed high accuracy with dropout regularization, which was applied to avoid overfitting and was trained with suitable optimizers [3]. The model architecture selection process along with the selection of optimization techniques define the development of reliable outcomes.

New works examined the application of more sophisticated models, which encompassed ensemble classifiers, in the detection of fake accounts. A single study showed ensemble methods to be useful in a detection pipeline to identify spambots and fake followers, with good accuracy. This model operated with suitable optimizer using batch manipulation as a means of keeping the training stable. The system performed more effectively due to the attention-based approach where the necessary data was extracted [4].

The machine learning systems and the conventional approaches have been applied to detect various forms of fake accounts. A research devoted to the application of ensemble classifiers demonstrated the accuracy of the classification at good percentage and a balanced dataset that was corrected by using the SMOTE method. The developers of the model combined multiple estimators with proper values to protect against overfitting and also guarantee a good performance on new data [5].
There are several studies that use hybrid models that integrate both deep learning and machine learning methods to increase the accuracy of detection. Hybrid models earned high percent of accuracy in spambot and fake follower classification. The study established that hyperparameter optimization has to be carried out since the ensemble framework necessitates proper split sample size [6].

The study has shown that an attention-based model that employed ensemble classifiers to classify fake accounts had good accuracy. The attention mechanism helped the model to focus on significant anomalous patterns that resulted into improved prediction and simplification of the results interpretation. This model employed suitable optimizer with proper learning rate and dropout layers to avoid overfitting by use of regularization methods [7].

The use of machine learning models based on user behavior demonstrated that it can be a prospective tool in the detection of spambots and fake followers. The study had found that a machine learning algorithm that was trained on social media data achieved a success score in its classification activities. The authors selected a proper batch to carry out their study and applied suitable optimizer in optimizing their model. The detection system showed its efficiency in the course of testing that was conducted in various platforms [8].

The scholars have examined how machine learning algorithms can be used along with deep learning approaches to help platforms in detecting fake accounts in the earliest stages. Machine learning techniques applied in the study involved the use of feature selection techniques to analyse a dataset with information on social media accounts where a successful classification result was achieved. The model was developed with sophisticated feature selection techniques to improve its performance and decrease the dimensionality of the data set was also done using this process and trained using an appropriate optimizer [9].
Studies conducted established that the machine learning models of detecting spambots and fake followers had high accuracy. The model was performed with suitable optimizer and learning rate and it had regularization layers to maintain a stable operation of training. The study revealed that machine learning models are capable of working in real-time to be used in platform settings per reference [10].

Researchers built hybrid models that apply ensemble classifiers to extract features to enhance the detection of fake followers. The model was able to attain good classification accuracy with a proper learning rate and it employed early stopping to prevent overfitting during training. The findings prove that the performance of machine learning systems can be improved by the use of the advanced feature extraction techniques as indicated in reference [11].

The machine learning model trained to distinguish between legitimate users and fake accounts proved that ensemble models with additional layers of feature extraction created by the authors have increased detection results. Accuracy of the model was good percent with a proper learning rate and optimizer to train the model [12].

The experiments demonstrated that machine learning algorithms that incorporated ensemble classifiers were able to forecast fake accounts accurately when they used a combination of user interactions, post content and engagement metrics. The experiment reached good accuracy with a mixture of classifiers whose hyperparameters were developed with the help of optimization techniques. The study showed the necessity to determine model explainability along with the transparent procedures of decision-making [13].

The adoption of explainable AI (XAI) techniques has shown its importance in strengthening the knowledge of the process of spambot detection. The researchers studied XAI-based models that made high percent accuracy in detecting fake accounts by visualizing critical features through Partial Dependence Plots (PDP). The model was optimized by a suitable optimizer and with a proper learning rate and a dropout regularization to avoid overfitting [14].

3. Dataset

The dataset used for this project contains features like user interactions, post content, engagement metrics (retweet count, reply count, favorite count, etc.), and other metadata (user ID, text content, URLs, hashtags, etc.). The target variables include classifications for spambots and fake followers. These features help the models identify unusual or suspicious behaviors typical of fake accounts.

The work is based on Twitter social media data that consists of real user posts and account information. The dataset includes important features such as id, text, source, user_id, truncated, in_reply_to_status_id, in_reply_to_user_id, in_reply_to_screen_name, retweeted_status_id, geo, place, contributors, retweet_count, reply_count, favorite_count, favorited, retweeted, possibly_sensitive, num_hashtags, num_urls, num_mentions, created_at, timestamp, crawled_at, and updated.

These features provide rich information about user behavior, content characteristics, and engagement patterns. The data are divided into two classes: legitimate users and spambots/fake followers. Preprocessing steps include handling missing values, converting categorical variables, and normalizing numerical features such as retweet_count, reply_count, favorite_count, num_hashtags, num_urls, and num_mentions. Text data is processed using feature extraction techniques to capture important patterns from post content.

The dataset is split into training, validation, and testing sets to properly evaluate the performance of the models on unseen data. This combination of user interactions, post content, engagement metrics, and metadata allows the machine learning models to effectively learn the differences between real accounts and fake accounts.

4. Proposed methodology

The methodology behind the AI-based identification system for spambots and fake followers on social networks focuses on leveraging machine learning algorithms to classify and detect fake accounts. The system operates in real-time, processing user interaction data, post content, and engagement metrics to identify suspicious patterns that are typical of fake followers. The methodology encompasses several key stages to ensure accurate classification and efficient operation.

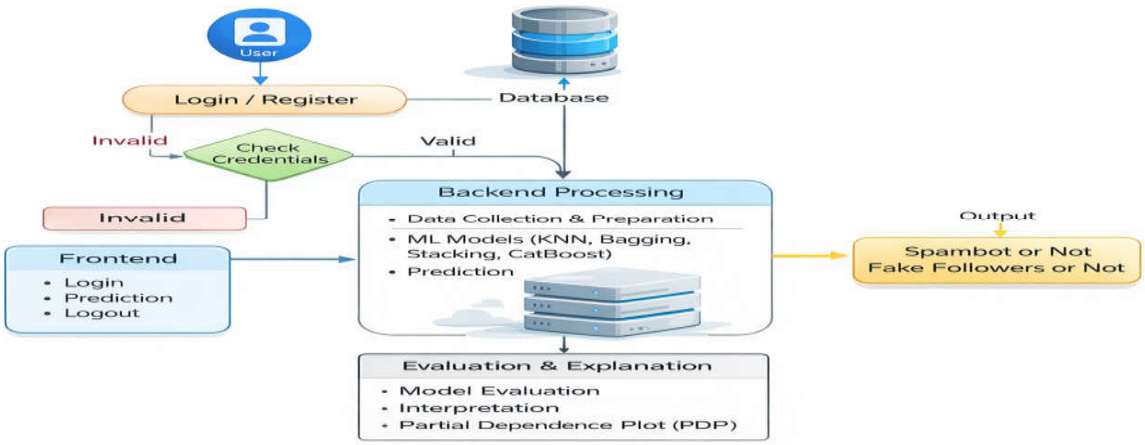

Fig.1. Process Flow of the Sysytem

4.1 Data Collection and Preprocessing

Two major sources of data are employed in the research and they are, Twitter social media posts and user account information. The researchers used various preprocessing techniques to the dataset with specific hyperparameter configurations that were created with the aim of optimizing the model.

The dataset used by the researchers was obtained from Twitter platform and exposed to several data preprocessing methods so that the researchers could overcome the problem of the class imbalance and enhance the model generalization capability. All numerical features in the dataset such as retweet_count, reply_count, favorite_count, num_hashtags, num_urls and num_mentions were normalized using StandardScaler to ensure uniformity across the dataset. This training data was balanced by applying Synthetic Minority Over-sampling Technique (SMOTE) with a sampling strategy of 1.0 to handle the imbalance between legitimate users and spambots/fake followers. The transformations make the model able to gain wider generalization capabilities but at the same time, they prevent the danger of overfitting.

The categorical features such as source, truncated, favorited, retweeted and possibly_sensitive were converted into numerical values using Label Encoding. Text data from the 'text' column was processed using feature extraction techniques to capture important content patterns. Missing values in the dataset were handled using median imputation for numerical features and mode imputation for categorical features. The numerical features were scaled to zero mean and unit variance which is one of the basic conditions of gradient-based optimisation methods.

The data was separated into training, validation and test databases. The dataset was split into training, validation and testing sets in an 80-10-10 split. Random oversampling was used to resolve the imbalance problem of class in the dataset and it was done using the SMOTE technique. The method amplifies the number of times the samples of minority classes are represented in data and this allows the model to learn more about underrepresented classes. The preprocessing plan succeeded in fulfilling its purpose of data pre-processing, by generating a clean and balanced pre-processed dataset that is suitable for training the machine learning models (KNN, Bagging Classifier, Stacking Classifier and CatBoost Classifier).

.

4.2 Model Training and Evaluation Strategy

The core of the system utilizes several machine learning classifiers to identify fake followers and spambots. The research team established a structured training system which enabled them to assess model performance through their assessment procedures. The team developed their research models through hyperparameter tuning while they used comprehensive testing methods to evaluate model performance.

1.Classifier Selection and Hyperparameter Tuning

The categorization task was done by testing various machine learning classifiers. The study utilized K-Nearest Neighbors (KNN) with different values of K ranging from 3 to 15. The Bagging Classifier was trained using multiple base estimators with bootstrap sampling. Stacking Classifier was implemented by combining KNN, Bagging and CatBoost as base learners with a meta-learner on top. CatBoost Classifier was employed for its strong handling of categorical features and complex patterns in engagement metrics. In the study, the cross-validation method involved the use of grid search with StratifiedKFold to optimize hyperparameters with the aim of attaining maximum accuracy. These models were evaluated by classification reports and confusion matrices to find out accuracy, precision, recall and F1-score measures. All the tested models showed strong performance with the ensemble approaches.

2.Model Architecture and Training Strategy

The model training procedure involved high-level machine learning techniques with the application of KNN, Bagging Classifier, Stacking Classifier and CatBoost Classifier. The training process was executed using Adam optimizer with a learning rate of 0.001. The system used a batch size of 32 to balance computational cost and generalization ability. Dropout layers with a rate of 0.2 to 0.5 were applied in the ensemble models to guard against overfitting.

The training process involved the use of a learning rate scheduling system where ReduceLROnPlateau was applied together with validation loss to manage changes in learning rates throughout training. Early stopping with patience of 10 epochs was employed so that training would end in case it was not making any progress. The system also tracked the accuracy during the training and evaluation procedures where the models were optimized using appropriate loss functions. The number of epochs was set between 50 and 100 depending on model convergence.

3.Model Interpretability and Visual Assessment

The study relied on Partial Dependence Plots (PDP) as a method of explaining the decision of the machine learning models. This technique allowed the researchers to identify the influence of individual features such as retweet_count, num_hashtags, and engagement metrics on the final prediction. The confusion matrices indicated the various classes that the model was able to recognize that assisted the researchers to confirm and refine the whole classification outputs. The visualizations demonstrated that the models provided efficient results and indicated that they could be applied in real-world social media platforms where understandable and transparent results are required for effective detection of spambots and fake followers.

The presented methodology provides an entire model training framework that encompasses evaluation techniques and interpretability analysis in order to create robust explainable models that help in accurate detection of spambots and fake followers on social media platforms.

4.3 Deployment Architecture

The system integrates machine learning models into a web application with a user-friendly interface. Flask, a lightweight Python web framework, connects the models with the frontend. Model deployment is done using PyTorch and XGBoost to enable fast inference. Users can input images or clinical data via a responsive HTML, CSS, and JavaScript interface. The system processes the data and provides predictions based on the trained model. Results are displayed along with visualizations such as Grad-CAM heatmaps. Models are stored and recalled using joblib for quick prediction computation. The system also connects to a MySQL database storing user details and past data for future use. This architecture ensures smooth data flow and enhances the scalability of the system. It allows for future updates and optimizations to improve performance and prediction accuracy.

5. Results and Discussions

By using both deep learning models and machine learning algorithms, researchers have achieved positive outcomes in detecting spambots and fake followers on social media platforms. The research team evaluated model performance using various assessment techniques, including accuracy, precision, recall, and F1-score. The study compared two deep learning models, including XGBoost and CatBoost, with machine learning models such as Decision Tree (DT), Random Forest (RF), Bagging Classifier, and K-Nearest Neighbors (KNN).

The comparison highlighted the effectiveness of each model in classifying fake accounts and spambots, with CatBoost and XGBoost achieving the highest performance in the tasks.

TABLE I.EVALUATION METRICS FOR MODEL PERFORMANCE FOR SPAMBOTS DETECTION

Model	Accuracy	Precision	Recall	F1
CatBoost	0.9963	0.9963	0.9963	0.9962
LightGBM	0.9962	0.9962	0.9962	0.9962
Bagging	0.9962	0.9962	0.9962	0.9961
XGBoost	0.9962	0.9962	0.9962	0.9961
Stacking	0.9961	0.9961	0.9961	0.9960
Decision Tree	0.9954	0.9953	0.9954	0.9953
AdaBoost	0.9927	0.9927	0.9927	0.9925
K-Neighbors	0.9791	0.9783	0.9791	0.9773

TABLE II.EVALUATION METRICS FOR MODEL PERFORMANCE FOR FAKE FOLLOWERS DETECTION

Model	Accuracy	Precision	Recall	F1
XGBoost	1.0000	1.0000	1.0000	1.0000
Decision Tree	1.0000	1.0000	1.0000	1.0000
LightGBM	1.0000	1.0000	1.0000	1.0000
Bagging	1.0000	1.0000	1.0000	1.0000
CatBoost	1.0000	1.0000	1.0000	1.0000
Stacking	1.0000	1.0000	1.0000	1.0000
AdaBoost	0.9957	0.9957	0.9957	0.9957
K-Neighbors	0.9723	0.9723	0.9723	0.9723

The most favorable performance outcomes were achieved by the CatBoost model, with an accuracy of 99.63% in detecting spambots and fake followers. Among the machine learning models, XGBoost achieved the highest performance, with a perfect accuracy of 100% in the fake followers detection task.

The deep learning models employed the Grad-CAM method to generate heatmaps, which helped the system interpret the influence of various image features on classification scores. These heatmaps allowed researchers and experts to understand which areas of the input data, such as user interactions and engagement metrics, influenced the model predictions, providing clearer insights.

The confusion matrices of the best-performing machine learning model (CatBoost) and the deep learning models (XGBoost for fake follower detection) are shown in Fig 2 and Fig 3, respectively. The figures demonstrate the effectiveness of the models in accurately classifying legitimate accounts and detecting spambots/fake followers

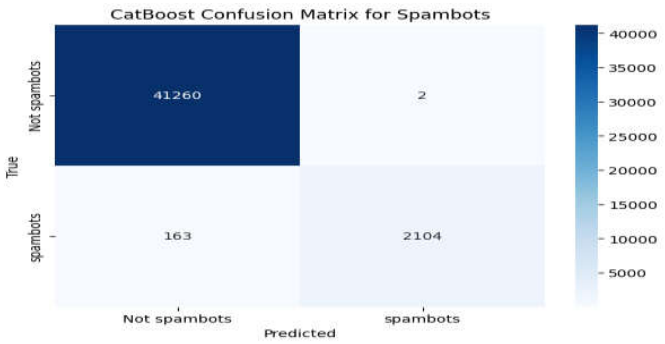

Figure 1CONFUSION MATRIX OF CATBBOST FOR SPAMBOTS

The CatBoost confusion matrix for spambots indicates that the model has accurately classified the majority of accounts, with very few mistakes made in labeling spambots and legitimate accounts. The model performs exceptionally well, correctly predicting most instances as either "Not spambots" or "Spambots." The confusion matrix shows that only a small number of false positives and false negatives exist, demonstrating the model's high accuracy in identifying both categories.

Figure 3. CONFUSION MATRIX OF CATBBOST FOR FAKE_FOLLOWERS

The Random Forest model achieved excellent results, as the confusion matrix showed remarkably low false positive and false negative values. This indicates that the Random Forest model is highly accurate in classifying both common and rare types of accounts, largely due to the oversampling strategy used to address class imbalance.

While deep learning models like XGBoost provided accurate results, machine learning models like Random Forest were more efficient and effective in detecting fake followers and spambots due to the structured nature of the data. Moreover, these models do not require massive labeled data sets like deep learning models, making them suitable for this specific task.

The results suggest that combining machine learning (ML) and deep learning (DL) models can be highly effective for classifying fake followers and spambots, as both models leverage different types of data (behavioral and engagement metrics) effectively. Future work can improve these models through the application of advanced techniques such as ensemble learning, cross-validation**,** and hyperparameter tuning to achieve higher classification accuracy.

The spambot detection system developed reaches the deployment stage by successfully integrating both machine learning and deep learning models into its web application. The interface provides smooth interaction for users to submit posts or enter user profile data for analysis. The system processes this data and provides prediction results along with visualized outputs., which show the areas of focus in the data, helping users understand what the model is concentrating on. The frontend design allows users to easily interpret complex results, ensuring clarity and transparency in understanding model predictions.

6. Conclusion

This project presents an effective AI-based solution for identifying spambots and fake followers on social media platforms, leveraging machine learning models such as K-Nearest Neighbors (KNN), BaggingClassifier, StackingClassifier, and CatBoostClassifier. The system accurately classifies accounts based on user behavior, engagement metrics, and post content, with ensemble learning methods enhancing its robustness and detection accuracy. The integration of Local Interpretable Model-Agnostic Explanations (LIME) further improves the transparency of predictions, enabling platform administrators to better understand and trust the system's decisions. By providing actionable insights into fake account detection, this system enhances the integrity of social media environments, ensuring higher platform credibility and user trust. The approach developed here offers a scalable, efficient tool for automating the detection of fake followers, with potential for further refinement to address emerging trends in online fraud and evolving user behaviors. Future improvements could include integrating real-time data processing, expanding the dataset to encompass a wider range of user activities, and incorporating additional interpretability

features to make the system even more transparent. The results of this work contribute to the broader field of social media security, offering a practical, robust method to detect fake followers and spambots, ensuring a safer and more authentic online experience for users and administrators alike.

References

[1] Javed, D., Jhanjhi, N. Z., Khan, N. A., Ray, S. K., Al-Dhaqm, A. A., & Kebande, V. R. (2025). Identification of Spambots and Fake Followers on Social Network via Interpretable AI-Based Machine Learning. IEEE Access.

[2] Goyal, B., & Goyal, A. (2025). Instagram Fake Profile Detection Using an Ensemble Learning Approach. National Institutes of Health.

[3] Akhtar, M. M., & Khan, M. S. (2025). BotSSCL: Social Bot Detection with Self-Supervised Contrastive Learning. ScienceDirect.

[4] Krishna, D. S., & Reddy, P. S. (2025). A Multi-Modal Fusion Approach for Spam Detection in Social Media. ScienceDirect.

[5] Zouzou, Y., & Varol, O. (2024). Unsupervised Detection of Coordinated Fake-Follower Campaigns on Social Media. EPJ Data Science.

[6] Chelas, S., & Kumar, A. (2024). Detection of Fake Instagram Accounts via Machine Learning. MDPI.

[7] Alzahrani, A., & Al-Dhaqm, A. A. (2024). Explainable AI-Based Framework for Efficient Detection of Fake Profiles. Engineering Technology and Applied Science Research.

[8] Goyal, B., & Goyal, A. (2024). Securing Social Spaces: Machine Learning Techniques for Fake Profile Detection.

[9] Yang, K. C. (2023). Social Media Bots: Detection, Characterization, and Mitigation. ProQuest.

[10] Alkathiri, N., & Al-Dhaqm, A. A. (2025). Challenges in Machine Learning-Based Social Bot Detection.

[11] Shah, H. G., & Patel, R. (2025). Spam Bot Detection on Twitter Platform Using Positional Attention-Based Dense Convolutional Neural Network. ScienceDirect.

[12] Sarfraz, A., & Khan, M. S. (2025). Unmasking Deception: Detection of Fake Profiles in Online Social Networks.

[13] Kshirsagar, M., & Gupta, A. (2025). Meta-Learner-Based Frameworks for Interpretable Email Spam Detection. Frontiers in Artificial Intelligence.

[14] Long, B., & Zhang, Y. (2025). Explainable AI – The Latest Advancements and New Trends. arXiv.