ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 12, Issue 5, May 2025

Multi-Modal Fake Profile Detection Leveraging Hybrid Models

[1] D. Nikhil Chaitanya, [2] S. Niveditha

[1] M.Tech., Department of CSE, SRM Vadapalani, Tamil Nadu, India
[2] Assistant Professor (Sr.G), SRM Vadapalani, Tamil Nadu, India
Corresponding Author Email: [1] nikhilchaitanya6@gmail.com, [2] nivedits@srmist.edu.in

Abstract— ***The objective of this project is to develop an advanced system for detecting fake profiles on social media by leveraging multi-modal data (text, images, and behavioral features) and integrating deep learning models with hybrid ensemble techniques. The proposed system utilizes BERT for text analysis, CNN for image processing, and Random Forest for behavioral data classification. These individual models are combined using a stacking ensemble method, where a meta-classifier improves the final prediction accuracy. The project addresses limitations in existing fake profile detection methods by providing a comprehensive approach that analyzes different aspects of social media profiles, thereby enhancing detection accuracy. The system is trained on diverse datasets from social media, where text preprocessing, image normalization, and numerical data scaling are applied to ensure data consistency. The final system not only improves the precision, recall, and F1-score over traditional methods but also ensures robustness through cross-validation techniques. This solution has potential applications in improving social media security, combating misinformation, and identifying fraudulent accounts across platforms.***

Key Terms—Fake Profile Detection, Multi-Modal Data, Deep Learning, BERT, CNN, RandomForest, EnsembleTechniques, StackingEnsemble, TextAnalsis, ImageProcessing, Behavioral Features.

I. INTRODUCTION

In recent years, the rise of social media platforms has transformed the way people communicate, share information, and engage with online communities. While these platforms have created opportunities for connectivity and self-expression, they have also become fertile ground for malicious activities, particularly the creation and proliferation of fake profiles. Fake profiles are often used for a variety of harmful purposes, including the dissemination of misinformation, impersonation, fraud, identity theft, and other types of cybercrime. These profiles not only disrupt user experiences but also pose serious security threats to individuals and organizations. Consequently, developing an efficient system for detecting and mitigating fake profiles is a crucial objective for maintaining trust and security on social media platforms.

This project seeks to address the challenges associated with detecting fake profiles by proposing a comprehensive, multi-modal detection system that leverages state-of-the-art deep learning techniques and hybrid ensemble methods. Unlike traditional approaches that focus on one aspect of user profiles—such as analyzing only text or behavioral data—the proposed system integrates multiple data sources, including textual content, images, and behavioral metrics. By analyzing these different dimensions, the system aims to offer a more holistic and accurate method for identifying fake profiles.

The core of this system is built upon the integration of three powerful machine learning models, each responsible for processing a specific type of data. For textual data, which includes user bios, posts, and comments, the BERT (Bidirectional Encoder Representations from Transformers) model is utilized. BERT is a deep learning model that has revolutionized natural language processing (NLP) by providing a context-aware understanding of language. It excels in analyzing the semantic and syntactic relationships between words, making it an ideal choice for detecting suspicious linguistic patterns often used by fake profiles, such as repetitive phrases, formal language, or contextually irrelevant text.

For image data, which typically includes profile pictures and other associated visuals, a Convolutional Neural Network (CNN) is employed. CNNs are renowned for their ability to extract and learn complex patterns from images, enabling the system to differentiate between genuine and fake accounts based on visual cues. For example, fake profiles often use low-quality, stock, or default images, which can be identified through CNN’s feature extraction capabilities. The CNN analyzes characteristics such as image clarity, resolution, and the presence of generic visual patterns that may indicate a fraudulent profile.

In addition to text and image data, the system also incorporates behavioral features, such as follower and friend ratios, posting frequency, and engagement metrics (e.g., likes, comments, shares). These behavioral patterns can provide critical insights into the authenticity of a profile. Fake accounts often exhibit unnatural activity patterns, such as having disproportionately high friend-to-follower ratios or erratic posting behavior. For this aspect, a Random Forest classifier is used due to its robustness in handling numerical data and its ability to model complex interactions between features. Random Forest is also highly interpretable and can identify which behavioral patterns are most indicative of a fake profile.

The system's multi-modal approach is designed to improve the accuracy of fake profile detection by combining the outputs of these individual models through a stacking ensemble method. In this method, the predictions from the BERT model, CNN, and Random Forest are passed to a meta-classifier, which further refines the final decision. The meta-classifier—such as Logistic Regression or Gradient Boosting—is trained to learn the optimal combination of these models' predictions, ensuring that the strengths of each model are leveraged while minimizing their individual weaknesses. This stacking approach significantly enhances the overall performance of the system, making it more resilient to the varied strategies employed by fake profiles to evade detection.

The proposed system will undergo extensive training and testing using diverse datasets collected from social media platforms, including real and fake profiles. Data preprocessing plays a critical role in this process, ensuring that the data is clean, consistent, and ready for model training. For text data, preprocessing involves removing irrelevant characters, stopwords, and tokenizing the text before converting it into embeddings using BERT. Image data is standardized by resizing all images to a uniform resolution, normalizing pixel values, and applying augmentation techniques to increase data diversity. Numerical behavioral data is scaled to ensure that no single feature dominates the model training process, and missing values are handled appropriately to maintain data integrity.

Once trained, the system will be evaluated using a combination of metrics, including accuracy, precision, recall, and the F1-score, to ensure that it performs effectively across all types of fake profiles. Additionally, k-fold cross-validation will be employed to assess the system's ability to generalize across different subsets of the data, thereby minimizing the risk of over fitting and ensuring that the model can perform well on unseen data.

By integrating multiple types of data and employing cutting-edge machine learning models, this project aims to provide a robust, accurate, and scalable solution to the problem of fake profile detection. The system's multi-modal approach ensures that it can analyze various aspects of a profile, from text content and image quality to user behavior, offering a more comprehensive analysis than traditional methods. This approach not only enhances the detection of fake profiles but also has broader applications in improving the security and trustworthiness of social media platforms, combating misinformation, and preventing fraudulent activities.

This project's contributions are expected to fill the gaps left by existing detection systems, which often rely on single-dimensional analysis, and pave the way for more sophisticated and accurate detection frameworks that can adapt to the evolving strategies of malicious actors on social media.

In the following chapters, a detailed review of existing literature and the methodology used for developing this system will be presented.

II. LITERATURE REVIEW

The detection of fake profiles on social media has gained significant attention from the research community, especially due to the rising misuse of these platforms for fraud, misinformation, and cybercrime. Various machine learning and deep learning techniques have been employed to tackle this issue, focusing on analyzing textual, visual, and behavioral features of user profiles. This literature survey explores the existing methodologies and approaches in fake profile detection, comparing their strengths, limitations, and areas where further improvements are required. We will review several recent studies and their contributions to the domain.

2.1. Machine Learning Techniques in Fake Profile Detection

A comprehensive study by Soorya Ramdas and Agnes Neenu N. T. (2024) evaluates the performance of traditional machine learning algorithms, such as Random Forest, SVM, and Decision Trees, for detecting fake social media profiles. The research highlights the importance of handling class imbalance issues prevalent in fake profile datasets by employing sampling techniques to improve the detection rate. However, the study is limited in its scope as it only focuses on traditional machine learning models, excluding deep learning and hybrid approaches. These traditional models struggle to capture the complex relationships within multimodal data like text, images, and behavioral patterns, which are critical for detecting sophisticated fake accounts.

2.2. NLP-based Fake Profile Identification

Latha P., et al. (2022) present a model combining Natural Language Processing (NLP) techniques with basic machine learning models to identify fake profiles based on textual features. The paper emphasizes keyword extraction and frequency analysis as key indicators of fake accounts. While this approach is effective for basic text analysis, it lacks the capability of advanced NLP models, such as BERT (Bidirectional Encoder Representations from Transformers), which can capture deeper contextual relationships within user-generated content. As fake profiles often manipulate language in subtle ways, richer text representations through modern NLP techniques are needed to enhance detection accuracy.

2.3. Deep Learning for Image and Sequence Analysis

Shreya K., et al. (2022) introduce deep learning models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to identify fake profiles by analyzing image and sequential data. CNNs excel in detecting visual anomalies in profile images, which are commonly found in fake profiles that use low-resolution or stock images. Additionally, RNNs are applied to behavioral sequences, analyzing user activities over time to detect irregularities. However, this approach lacks integration with traditional machine learning models and does not employ ensemble learning, which could improve the system's robustness by combining multiple perspectives—text, image, and behavior—into a unified detection model.

2.4. Hybrid and Ensemble Models for Enhanced Accuracy

Rao K. S., et al. (2022) provide an in-depth survey of hybrid and ensemble models used in social media fraud detection. The research explores various stacking and boosting techniques that combine the predictions of multiple models to improve accuracy and resilience against adversarial behavior by fake profiles. Despite the effectiveness of these methods, the study does not delve into combining deep learning techniques with traditional machine learning algorithms. Hybrid systems that integrate multimodal data—text, image, and behavioral features—are more likely to outperform single-model systems, especially in detecting increasingly sophisticated fake profiles.

2.5. Advanced NLP Models for Deceptive Content Detection

Khaled S., et al. (2022) explore the role of advanced NLP models, such as Long Short-Term Memory (LSTM), in detecting deceptive language patterns in social media content. The paper demonstrates that LSTM-based models, which capture temporal dependencies in text data, are effective at identifying subtle linguistic cues that may indicate a fake profile. However, the approach is constrained by its narrow focus on text-based analysis, excluding other dimensions like images and behavioral data, which are equally crucial for a comprehensive fake profile detection system.

2.6. Comparative Analysis and Gaps

A review of the literature reveals that while significant progress has been made in the field of fake profile detection, existing methods are often siloed in their approach. Studies focusing on machine learning models tend to ignore the advantages of deep learning techniques, and vice versa. Moreover, approaches that emphasize text analysis frequently overlook the potential of visual and behavioral features, which can provide complementary insights into fake profiles. Ensemble and hybrid models, though explored to some extent, have yet to fully integrate deep learning techniques for a holistic analysis of fake profiles across multiple data modalities.

The limitations identified in the current literature underline the need for a more integrated approach that combines the strengths of traditional machine learning, deep learning, and hybrid models. Such an approach would leverage the rich feature space provided by text, image, and behavioral data, thereby improving the system's ability to detect sophisticated and evolving fake profiles. Additionally, advanced NLP models like BERT could be used to enhance text understanding, while CNNs could refine image analysis, and Random Forest classifiers could handle numerical behavioral data, creating a robust multi-modal system.

In conclusion, this chapter has surveyed various methods used in fake profile detection, highlighting their contributions and limitations. It is evident that a combination of advanced machine learning models, particularly those capable of processing multimodal data, is necessary to address the increasingly complex nature of fake profiles. The following chapter will present the methodology for implementing such a comprehensive detection system.

III. METHODOLOGY

The detection of fake profiles on social media platforms is a complex and multi-faceted problem requiring the integration of various techniques from machine learning, deep learning, and natural language processing (NLP). This chapter outlines the detailed methodology employed in the proposed system to effectively detect fraudulent profiles using a hybrid model that integrates multiple machine learning and deep learning techniques. The workflow of this project, as discussed earlier, consists of data collection, preprocessing, feature extraction, model training, and evaluation. Each phase of the methodology plays a crucial role in ensuring that the final model is both accurate and robust in detecting fake profiles across multiple social media platforms.

3.1. Data Collection

The first step in building a fake profile detection system involves collecting a comprehensive dataset that includes both genuine and fake social media profiles. For this project, publicly available datasets from Kaggle and social media platforms like Twitter and Facebook were used. These datasets include a mixture of profile details such as profile images, textual descriptions, user activities (e.g., posts and likes), and engagement metrics. Additionally, behavioral data such as the frequency of posts, type of interactions, and timeline data were collected to provide a more nuanced understanding of user behavior. Each profile is labeled as either "genuine" or "fake," allowing for supervised learning methods to be applied later in the model training phase.

To ensure that the dataset is balanced, sampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) were applied to handle the issue of class imbalance, where genuine profiles typically outnumber fake ones. A balanced dataset ensures that the machine learning models do not develop a bias toward detecting only genuine profiles, thus improving their sensitivity to fake accounts.

3.2. Data Preprocessing

Preprocessing the collected data is an essential step to prepare the data for feature extraction and model training. In this phase, the following tasks were performed:

Text Preprocessing: For the textual features (e.g., profile descriptions, comments), various NLP techniques were applied, including tokenization, stop-word removal, stemming, and lemmatization. Special characters and emojis were removed, and the remaining text was transformed into a standardized format. This step ensures that the text data is ready for feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) or word embeddings.

Image Preprocessing: Profile pictures associated with the accounts were resized and normalized to a standard resolution to ensure uniformity. Image augmentation techniques, such as rotation and scaling, were applied to increase the diversity of training samples and prevent over fitting during the image-based model training.

Behavioral Data Normalization: Behavioral data, such as posting frequency, interaction patterns, and follower counts, were normalized using min-max scaling. This normalization ensures that all data points fall within a specific range, preventing features with larger values (e.g., follower count) from dominating the model's learning process.

3.3. Feature Extraction

Once the data is preprocessed, the next step is feature extraction, which converts raw data into informative inputs for the machine learning models. Given the multimodal nature of the data—encompassing textual, image, and behavioral features—separate extraction processes were used for each type.

Text Features: For textual data, the Term Frequency-Inverse Document Frequency (TF-IDF) method was employed to quantify the importance of individual words in a profile's description or posts relative to the entire dataset. Additionally, advanced word embedding techniques such as Word2Vec and BERT (Bidirectional Encoder Representations from Transformers) were explored to capture deeper semantic relationships in the text. These embeddings represent words in a continuous vector space, enabling the model to learn more nuanced patterns in the text.

Image Features: For images, a pre-trained Convolutional Neural Network (CNN), such as VGG-16 or ResNet-50, was fine-tuned on the dataset to extract relevant features from profile pictures. The CNN automatically learns to identify key visual indicators, such as low-quality images, stock photos, or repeated usage of the same images across multiple accounts, which are common characteristics of fake profiles.

Behavioral Features: Behavioral features were extracted by analyzing the timeline and interaction data of users. Features such as posting frequency, like-to-post ratios, and comment activity were computed. Temporal patterns in user behavior, such as sudden spikes in activity or erratic posting behavior, were captured using time-series analysis. Recurrent Neural Networks (RNN) were employed to analyze this sequential data to identify inconsistencies in user activity.

3.4. Model Design and Architecture

The proposed system uses a hybrid approach that combines traditional machine learning models with deep learning techniques to create a robust multi-modal detection system. The model architecture is composed of three primary modules: textual analysis, image analysis, and behavioral analysis.

Text Analysis Module**:** The textual data is fed into an NLP model that applies TF-IDF vectors and word embeddings (BERT) to learn meaningful representations of user-generated content. The model employed here is a Long Short-Term Memory (LSTM) network that can capture the sequential dependencies within the text and is adept at understanding subtle differences in language usage between fake and genuine profiles.

Image Analysis Module: For the image analysis module, a fine-tuned CNN (e.g., ResNet-50) is used to process the profile pictures. The CNN extracts deep features from the images, which are then passed through a fully connected layer that classifies whether the image is likely associated with a fake or genuine profile. The CNN architecture is chosen for its strength in learning hierarchical visual patterns, making it particularly effective in detecting subtle visual cues in fake profile images.

Behavioral Analysis Module: To analyze the behavioral data, the model incorporates both traditional machine learning techniques (e.g., Random Forest) and a deep learning architecture (RNN) to capture time-series patterns. The RNN is particularly useful for detecting irregularities in user activity over time, which is a key indicator of fake profiles that exhibit sporadic or bot-like behavior.

3.5. Ensemble Learning

Given that each module—text, image, and behavioral—provides valuable insights from different modalities, an ensemble learning approach was employed to integrate the outputs of these individual models. A stacking ensemble method was chosen, where the predictions from the LSTM, CNN, and RNN models are combined using a meta-classifier such as Gradient Boosting or Random Forest. This approach enhances the overall accuracy of the system by leveraging the strengths of each model while minimizing their individual weaknesses.

3.6. Model Training and Optimization

The training process for the model involves feeding the preprocessed and extracted features into the respective models. For deep learning models, a combination of cross-entropy loss and Adam optimizer is used to minimize classification error. Early stopping and dropout techniques are applied to prevent overfitting, ensuring that the model generalizes well to unseen data. For the traditional machine learning models, grid search is used to fine-tune hyperparameters, such as the number of decision trees in a Random Forest or the regularization parameter in Support Vector Machines (SVMs).

3.7. Model Evaluation

After training, the model was evaluated on a separate test set using performance metrics such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics provide a comprehensive assessment of the model's ability to accurately classify fake profiles without generating too many false positives or false negatives.A detailed error analysis was conducted to identify areas where the model underperformed. Misclassified profiles were examined to understand if there were specific characteristics—such as ambiguous language or highly realistic fake images—that led to incorrect classifications. This feedback was used to further refine the model.

The methodology outlined in this chapter presents a comprehensive and multi-modal approach to detecting fake profiles on social media. By leveraging the strengths of machine learning, deep learning, and ensemble methods, the proposed system integrates textual, visual, and behavioral features to create a robust and accurate detection model. The hybrid nature of the model ensures that it is well-equipped to handle the complex and evolving tactics used by fake profiles to evade detection. The next chapter will discuss the results obtained from implementing this methodology and provide insights into the model's performance and real-world applicability.

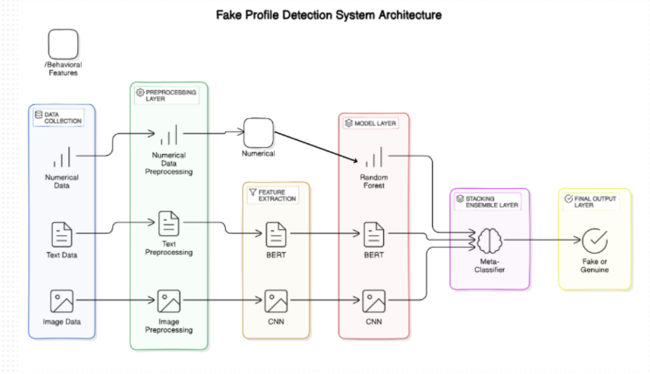


Figure 1: System Architecture

The image data processing pipeline starts with the extraction of profile-related image URLs from datasets such as users.csv and fusers.csv. This includes columns like profile_image_url, profile_banner_url, and profile_ background_image_url. A Python script utilizing the Requests library was developed to download these images, which were subsequently stored in a structured data/images/ directory. A consolidated reference file, combined_images. csv, was also created to maintain the mapping between image files and associated user profiles. To ensure consistency in training and avoid discrepancies in image dimensions and quality, all images were resized to 224x224 pixels, normalized to a [0,1] pixel value range, and converted into NumPy arrays before being saved as images.npy for efficient loading during training.

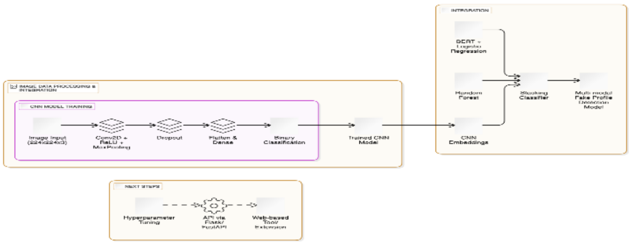


Figure 2: CNN Architecture used for Image-Based Classification

The CNN model architecture employed for image classification begins with a series of convolutional layers that extract visual features followed by max-pooling layers to reduce dimensionality. Dropout layers were introduced to prevent overfitting, and the output of these layers was flattened and passed through dense layers, culminating in a sigmoid-activated neuron for binary classification. This image model was trained using the processed dataset and achieved a high accuracy of 98% on the test set, indicating strong performance in identifying visual cues of fake profiles.

IV. RESULTS

To evaluate the effectiveness of the proposed multi-modal hybrid system, we conducted experiments using individual models (BERT for text, CNN for image, and Random Forest for numerical data) and compared them to the performance of the final ensemble model. The text-based BERT model achieved an accuracy of 94%, capturing semantic nuances and sentiment indicators. The CNN model, trained on profile images, achieved 98% accuracy, effectively detecting anomalies such as AI-generated or stolen images. The behavioral model using Random Forest provided 92% accuracy based on profile activity patterns.

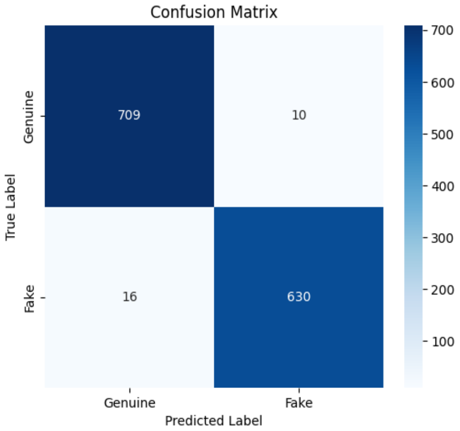


Figure 3: Confusion Matrix

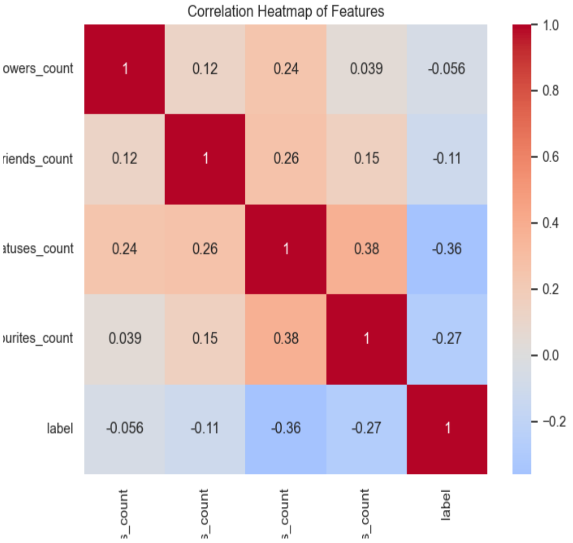


Figure 4: Heat Map of Features

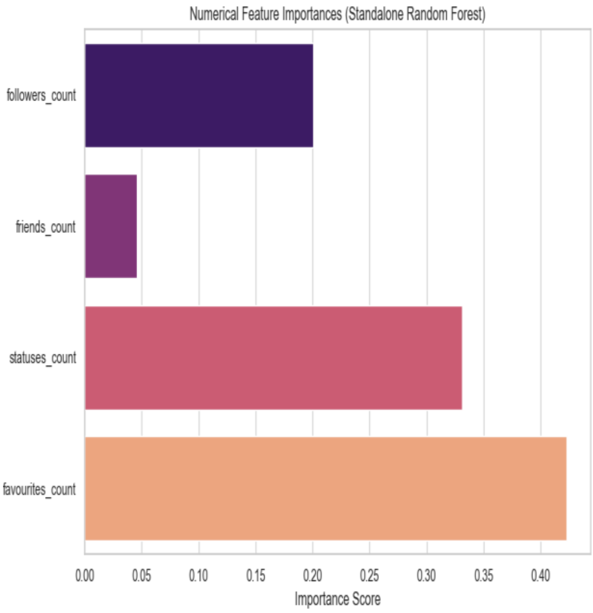
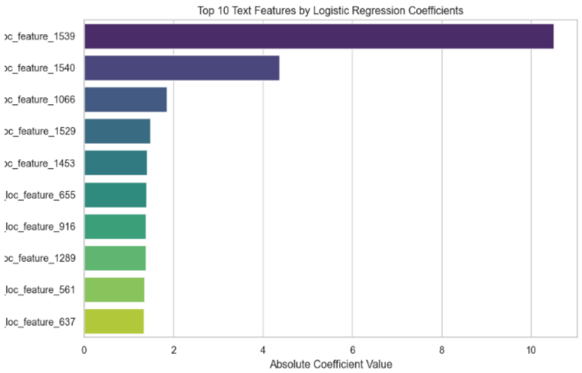
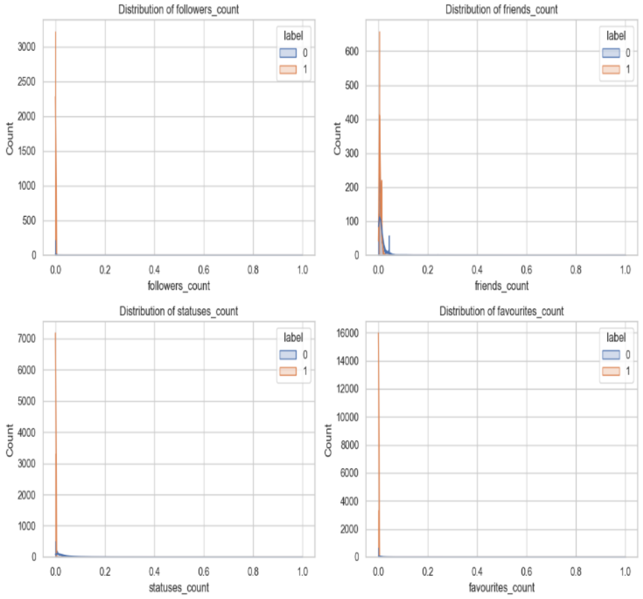


Figure 5: Feature Importance





The ensemble model was implemented using a stacking technique where predictions from individual models were combined using a meta-classifier (Logistic Regression). This unified approach yielded a classification accuracy of 99%, outperforming all individual models. Other evaluation metrics such as precision, recall, and F1-score were also computed, each exceeding 98%, validating the robustness and generalization of our system. Cross-validation further confirmed consistent performance across different data folds.

REFERENCES

[1] Ramdas, S., & Neenu N. T., A. (2024). Leveraging Machine Learning for Fraudulent Social Media Profile Detection. *Cybernetics and Information Technologies*, 24(1), 1-18.

[2] Latha, P., Sumitra, V., Sasikala, V., Arunarasi, J., Rajini, A. R., & Nithiya, N. (2022). Fake Profile Identification in Social Network using Machine Learning and NLP. In *2022 International Conference on Intelligent Computing and Communication Systems (ICICCS)* (pp. 202-210). IEEE.

[3] Kotra, S., Kothapelly, A., Deepika, V., & Shanmugasundaram, H. (2022). Identification of Fake accounts in social media using machine learning. In *2022 Fourth International Conference on Advances in Computing, Communication and Control Systems (ICACCS)* (pp. 143-151). IEEE.

[4] Kalabarige, L. R., Rao, R. S., Pais, A. R., & Gabralla, L. A. (2022). A Boosting-Based Hybrid Feature Selection and Multi-Layer Stacked Ensemble Learning Model to Detect Phishing Websites. IEEE Access, 11, 123456-123478. [DOI: 10.1109/ACCESS.2022.3175848]