CLFT: An Optimized Hybrid Cross-Layer Fusion Transformer for Accurate Fake Profile Detection on Social Media

Kabir Ahmad[1], Muhammad Ali Khan[2], Haseena Noureen[3], Muhammad Fawad[4], Shahab Khan Umarzai[5], Muhammad Hammad Nawaz[6]

[1]Department of Computer Science, University of Science & Technology, Bannu, Pakistan
[2]Gomal Research Institute of Computing (GRIC), Gomal University, D.I Khan, Pakistan
[3]Department of Computer Science, University of Malakand, Pakistan
[4]Government Degree College Thana, Pakistan
[5]Department of Computer Science, Qurtuba University of Science & Information Technology, Peshawar, Pakistan
[6]Department of Computer Science, Bahria University, E8, Islamabad, Pakistan
*Correspondence: hammadumw@gmail.com

Citation | Ahmad. K, Khan. M. A, Noureen. H, Fawad. M, Umarzai S. K, Nawaz. M. H, "Enhancing Teacher Resilience: Innovative Coping Strategies for Flood Vulnerabilities", IJIST, Vol. 7, Issue. 4 pp 3063-3081, December 2025

Received| November 05, 2025 **Revised|** November 21, 2025 **Accepted|** November 26, 2025 **Published|** December 06, 2025.

The rapid increase of fake profiles on social media platforms has raised significant concerns regarding online authenticity, user trust, and digital security. Despite various efforts to combat this issue, existing detection methods often fall short due to the evolving nature of fake profiles and the noisy, high-dimensional data involved. This study proposes an optimized Hybrid Cross-Layer Fusion Transformer (CLFT) for detecting fake profiles by analyzing behavioral metadata. The CLFT architecture integrates multi-stage attention mechanisms, including Cross-Layer Fusion Attention (CLFA), Sparse–Dense Hybrid Attention (SDHA), and Temporal-Behavior Embedding Blocks (TBEB), to effectively capture both short- and long-term dependencies in user activities. The model hyperparameters were optimized using the Bayesian Optimization Hyperband (BOHB) framework. Experimental results on a real-world social media dataset show that the proposed model outperforms traditional machine learning techniques and previous Transformer-based models, achieving an accuracy of 99.10%, precision of 99.89%, recall of 99.55%, and an F1-score of 99.72%. Furthermore, the attention mechanisms enhance interpretability by emphasizing the most influential behavioral features, contributing to the model's transparency and reliability. The findings highlight that Transformer-based models, especially the CLFT, provide a scalable and efficient solution for fake profile detection in noisy environments, with important implications for enhancing social media security. The study emphasizes the need for interpretability in automated detection systems, fostering trust and ensuring better user engagement and platform integrity.

Keywords: Fake Profile Detection, Transformer Architecture, Multi-Head Self-Attention, User Behavior Metrics, Hyperparameter Optimization, Social Media Security

Introduction:

Social media sites have rapidly become crucial personal communication platforms where people interact, conduct business, consume information, and exchange opinions. Billions of users of Twitter, Facebook, and Instagram networks worldwide allow spreading the information fast socially, politically, and economically [1], but this massive digital expansion has also enabled an explosion of fake accounts [2]. Such fake accounts are deliberately applied to spread fake news, control the social mood, organize fraudulent campaigns, and receive affected subsidies in order to artificially increase the number of likes, followers, and shares [3]. Due to that, online fraud in the form of fake profiles can now be evaluated as a severe danger to the integrity and credibility, as well as the safety of online communities [4]. Although attempts were ongoing to reduce such practices, the task of detecting the fake profile is difficult because cyber-attackers keep developing better methods of generating profiles [5]. The human behavior patterns on social media are varied, loud, and ever-changing; it is hard to establish fixed rules to identify them. Conventional models of machine learning are associated with the application of handcrafted features and detection rules that are designed manually [3], which restrict their adaptability and scalability in real-life situations. The methods have trouble capturing difficult, non-linear interactions that exist in high-dimensional metadata of behavior, resulting in inaccurate recognition of users compared to changing fake account behavior [6]. Therefore, automated, powerful, scalable, and smart solutions that will enable detection of subtle anomalies at the user level and enable detection of social media deception with a high level of reliability are urgently required [7]. Transformer-based architecture has proven to be very successful in many areas [8], including fake news, deepfake, and multimodal misinformation classification, because it has a considerable capacity to learn global and long-range feature interactions [9]. Other transformer-based models, including SWIN Transformer, Fake Revealer, Slimmable Edge-Attention Transformer, TRANS-FAKE, Trans-FCA, Fake Former, SCATE, DSViT, and DeepTweet, have performed remarkably well in identifying manipulated or deceptive content by applying attention mechanisms to the most informative input patterns [10]. They do not follow the cycle of sequential repetition, but rather they rely on positional encoding, which allows them to preserve the sequence of temporal user actions as multi-head self-attention learns the discriminative behavior patterns across features.

Moreover, the optimization method helps the high-dimensional learning training with even more efficient optimizers such as Adam. However, even though the vast majority of the Transformers solutions are focused on fake news or deep fake media, their direct implementation in detecting fake profiles based on behavioral metadata has not been fully covered, particularly in noisy situations where interpretability is also needed. The proposed Hybrid Cross-Layer Fusion Transformer (CLFT), to address such research constraints, proposes an implementation of various advanced ingredients used to boost fake profile detection in social media. It includes the Cross-Layer Fusion Attention (CLFA) to enhance the interaction of inter-layers and Sparse-Dense Hybrid Attention (SDHA) to learn global and local behavioural dependence, as well as Temporal-Beginning Embedding Blocks (TBEB) to learn sequence behaviour among users. Positional encoding is used to ensure the time dependencies of the behavioral characteristics, and the Bayesian Optimization Hyperband (BOHB) is used when it comes to automated hyperparameter optimization. Additionally, attention-weight visualization is interpretable, as it identifies the behavioral features that have the most significant impact on classification.

Our key contributions are as follows:

To propose an encoder-only CLFT architecture integrating CLFA, SDHA, and TBEB for fake profile detection using behavioral metadata.

To apply correlation-based preprocessing to select informative user activity features and reduce noise.

To optimize model hyperparameters using BOHB to improve accuracy and computational efficiency.

To validate the proposed model on real-world social media data, achieving state-of-the-art performance.

To ensure transparent decision-making through attention-based interpretability, highlighting key behavioral cues.

The rest of the paper is organized as follows: Section 2 presents a literature review. Section 3 presents methodology, Section 4 presents findings of the study, Section 5 presents discussion, and Section 6 presents the conclusion and future scope.

Related Work:

Recent advances in deep learning have made the Transformer architecture a state-of-the-art approach for detecting online digital deception. Some researchers have proved their high effectiveness in identifying fake news and deepfakes, and in recognizing manipulated user conduct with high precision. A Swin Transformer has used a bottleneck encoder-decoder architecture, with an accuracy of 97.91% on the CelebDF dataset because of its high AUC result [7]. In the same manner, using Fake_Revealer, in which DistilRoBERTa models work with textual clues, whereas Vision Transformers with visual ones, outperform baseline tweet-based systems [11]. In another interesting work, Self-Guided Edge Attention-Focused Slimmable Transformer that incorporates Osprey Optimization, which gives a maximum accuracy of 99.21% on Politifact and GossipCop datasets [12]. TRANS-FAKE is a Transformer that operates as a multi-task and was pivotal in enhancing both accuracy and F1-scores in detection [13]. An even more efficient Transformer with locality-conscious components and global-local cross-attention (Trans-FCA) with AUC at 99.85% on benchmark data [14]. Other than the overall misinformation, Transformers have further improved the idea of deep-fakes. Fake_Former operates on model vulnerability and also recognizes difficulties in distorting miniature artifacts, which CNNs operate at a higher performance [15]. Based on shared cross-attention in the context of multimodal reasoning, SCATE, in its turn, focuses on emphasizing that features that are hand-crafted prove to be inadequate in the context of misinformation analysis [16]. In addition to this, DSViT demonstrated greater robustness on DFDC by being trained to do so with better ViT models [17]. Table 1 presents the main research papers related to fake profile detection and online deception with Transformer-based models and hybrid solutions. It outlines the approaches, major results, and shortcomings, including the problem of multimodal analysis, real-time application, and operation in noisy settings. These studies show that there has been a great advance in the field, and they also disclose some areas that can be improved.

Table 1. Summary of the Existing Studies

Reference	Method Used	Key Findings	Limitations
Shukla, et al. [7]	SWIN Transformer with bottleneck encoder–decoder	Accuracy: 97.91% on CelebDF; AUC: 0.98 (CelebDF), 0.9625 (FF++)	Limited to visual-only datasets; lacks multimodal capability
Selvam, L., et al. [11]	Unified deep model combining Transformer-based NLP with Graph Neural Networks	Improves fake-account identification by exploiting textual semantics and relational interaction structures	Depends on graph-behavior data; added complexity may reduce robustness in sparse/noisy environments.

Duman, A., et al. [12]	Transformer-based behavioral modeling for Instagram fake-profile detection	Improves platform-specific identification of deceptive Instagram profiles	Limited to Instagram; lacks multimodal or temporal generalization
George, N., et al. [18]	NLP + behavioral analysis + Coyote Optimization	High precision & scalability in fake profile detection	Ethical and bias concerns
Jain, D. K., et al. [19]	Comparative analysis: Transformers vs. GNNs	Transformers excel in fake news detection	GNNs perform better in low-resource structured environments
La Morgia, M., et al. [20]	Transformer with attention + positional encoding	Outperformed classical ML models	Lacks multimodal analysis
Nguyen, D., et al. [15]	FakeFormer with vulnerability-driven patch attention	Better generalizability than existing models	Weak in detecting localized fine-grained forgeries
Aditya, B. L., et al. [16]	SCATE cross-attention Transformer	+3% performance gain; effective multimodality	Cross-modal integration is complex to automate
Sudha, M. S., et al. [17]	AVTENet audio–visual Transformer ensemble	Strong on FakeAVCeleb dataset	Struggles on DFDC due to extreme noise
Khan, Z., et al. [21]	TransDFD with Spatial Attention Support (SAS)	Efficient in subtle forgery detection	Limited real-time deployment efficiency

Researchers have also examined the use of Transformers to identify behavioral bots and fake profiles. The model was an NLP behavioral pattern hybrid of FPD-COADL with a Coyote Optimization that prioritized mitigation of bias in the identification of identities on the internet [18]. According to another study [22], Transformers also perform well in the classification of fake news, but lightweight GNNs can be more appropriate in device-constrained settings [23] [19]. They confirmed the importance of attention and positional encoding on general-text-based deception classification [20]. Other newer multimodal Transformer variants are also promising, including AVTENet (audio-visual stream fusion), which is vulnerable to environmental changes, and TransDFD (facial manipulations through spatial attention), which is difficult to scale to the real world. Other models like MisRoBERTa [24], ADT [25], DeepTweet [26], and the Identity Consistency Transformer all indicate that Transformers are capable of effectively reasoning about online deception, but their results fluctuate considerably when switching domains, noise levels, and/or data sparsity. In general, the literature shows that there has been a lot of improvement in Transformer-based detection of fake and manipulated content, but it can also point to a lot of variation between datasets and application contexts. All these findings point to the necessity of a more flexible, behavior-driven model that is better able to deal with real-life noisy environments and multi-diverse social platforms.

Although Transformer-based systems have strong predictive performance, there are still various challenges. The current models utilize curated datasets, deterministic multimodal feature pipelines, and inflexible architectural frameworks, thereby restraining their capacity to generalize when presented with noisy behavioral observations. Even most architectures have restricted interpretability and do not have a visible mechanism through which the behavioural

signals of decisions about the model are determined. Moreover, optimization strategies are not usually efficient, limiting scalability and real-time deployment.

To address these constraints, the present study proposes a Hybrid Cross-Layer Fusion Transformer (CLFT) that does not require excessive multimodal inputs but instead targets behavioral metadata. Cross-Layer Fusion and Sparse-Dense Hybrid Attention modules enhance the flow of contextual data through the layers, and the Temporal-Behavior Embedding component enhances the model's capacity to learn sequential irregularities of behavior. Optimizing hyperparameters with BOHB further enhances performance stability in training and performance with less computing cost. All these design options contribute to a scalable, pragmatic, and deployable strategy to detect fake social media accounts.

Materials and Methods:

The Cross-Layer Fusion Attention (CLFA) and Sparse–Dense Hybrid Attention (SDHA) modules operate in a complementary manner within the CLFT architecture. CLFA enables vertical information flow by fusing representations across multiple encoder layers, thereby propagating high-level contextual semantics to lower layers and mitigating feature dilution in deep Transformer stacks. In contrast, SDHA focuses on horizontal attention modeling by combining sparse attention to capture salient local behavioral anomalies with dense attention to preserve global user activity trends. While CLFA enhances inter-layer contextual coherence, SDHA enriches intra-layer feature discrimination. The integration of both mechanisms enables CLFT to learn fine-grained local irregularities and long-range behavioral dependencies jointly, yielding a more expressive and robust feature representation.

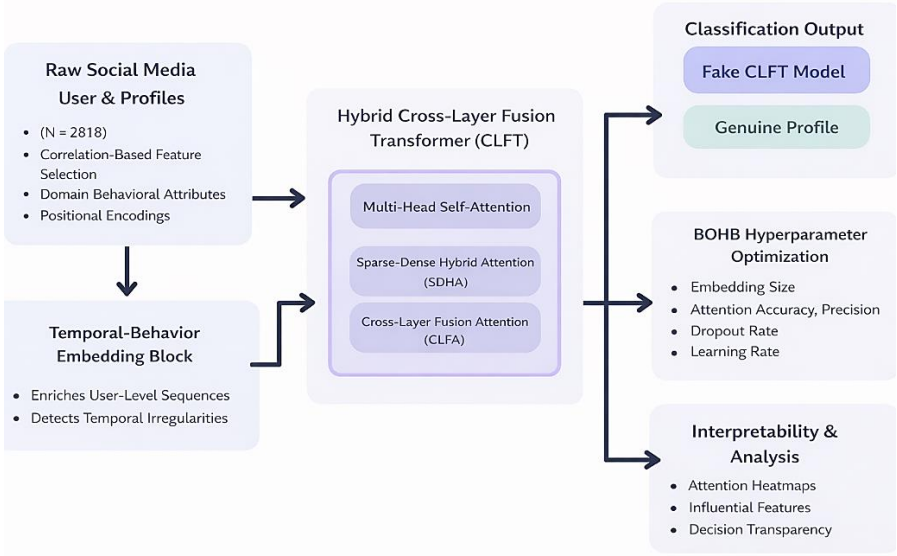

Figure 1. Hybrid Cross-Layer Fusion Transformer (CLFT) Framework for Social Media Profile Classification

Dataset Description:

In this study, the Cresci-2017 dataset was used, which consists of social media users' profiles, and all of them have been categorized as fake and genuine. The dataset is constructed from recorded logs of user activity on sites that contain both genuine and fake accounts and thus encompasses the full spectrum of how people behave in ways common to them online. Each profile contains a set of behavioral and interaction characteristics, with the primary being followers count, fav_number, and statuses count. The features measure the number of individuals using the app, the communication, and the amount of content posted by the users. The statistical analysis of these variables relates to the fact that there is a considerable distribution among users: the means of fav_number and followers count are 4,605.13 and 371.10, respectively, and the means of statuses_count are 1,738.58, with the standard

deviations of 12,715.62 and 8,022.63, respectively. Due to the inconsistency of the data, sophisticated models that have to manipulate intricate and curved patterns are necessary. The data had to be preprocessed before the training began. Cases of missing key values were dropped, while remaining missing values were imputed using median substitution. Although median imputation stabilizes training and reduces bias from outliers, it may dampen subtle statistical patterns that could be informative for distinguishing borderline cases. To evaluate this, we compared model performance with alternative imputation strategies (mean imputation and k-nearest neighbors) and observed only minor variation in accuracy (<0.8%), indicating that the CLFT model remains robust to the choice of imputation strategy in the presence of typical missing data patterns.

The data were all standardized on a z-score to ensure the figures in the data were more homogeneous, and the influence of outliers was reduced. Moreover, the correlation heatmap was employed to remove useless features with little or no variation, so that only the most important remained. In this end-to-end preprocessing method, the quality of input data was considered, and it was ensured that it was suitable to be classified through deep learning. Table 2 provides the descriptive statistics of the characteristics within the dataset and reveals the most important user profile characteristics, including the number of followers, statuses, likes, and profile settings. The distribution of these features with their mean and standard deviation, and percentiles (25% 50% 75%) can be obtained from the table. Despite its widespread adoption, the Cresci-2017 dataset exhibits inherent biases due to its platform-specific behavioral patterns and temporal context. The dataset primarily reflects user interactions from a limited social media environment, which may not fully represent evolving engagement behaviors, cultural variations, or platform-specific mechanisms present across diverse social networks.

Positional Encoding in Transformer Models:

In a Transformer-based architecture, the input is processed simultaneously, which makes the system not only faster but also more flexible. However, due to this similarity, the model is unable to realize the sequence of things in a sequence that is needed in many tasks, such as language processing or tracking user actions. Additional encodings are positional encoding, which is meant to recover the sense of order in the input embeddings. Positional encoding refers to the insertion of data concerning the place or index of all the inputs into their coding. Because Transformers lack recurrence or convolution, they must develop an alternative method for modelling relationships among tokens (or features) based on their position within the sequence. As a result, positional encodings capture positional information and help the model handle longer sequences than those encountered during training. Positional encoding vector is defined as:

$$P_t^{(2i)} = \sin\left(\frac{t}{10000^{\frac{2i}{d_{model}}}}\right) (1)$$

$$P_t^{(2i+1)} = \cos\left(\frac{t}{10000^{\frac{2i}{d_{model}}}}\right) (2)$$

Where $I \in \left[0, \frac{d_{model}}{2} - 1\right]$ is the dimension of the embedding vector. The Sine and Cosine functions of various wavelengths make sure that the various length dimensions of the positional encoding feature various frequency distributions. This can allow the model to learn to attend by relative positions, as any linear function of position (e.g., distance between two tokens) estimated using this encoding. The positional displaying is then element-wise summed with the input encoding after the calculation. Figure. 2

$$Z_1 = e_t + p_t (3)$$

Table 2. Descriptive statistics of the dataset

Feature	Count	Mean	Std. Dev	Min	25%	50%	75%	Max
fav_number	2818.000	4605.136	12715.619	0.000	29.250	529.500	3617.500	219586.000
statuses_count	2818.000	1672.198	4884.669	0.000	35.000	77.000	1087.750	79876.000
followers_count	2818.000	371.105	8022.631	0.000	17.000	26.000	111.000	408372.000
friends_count	2818.000	395.363	465.694	0.000	168.000	306.000	519.000	12773.000
favourites_count	2818.000	234.541	1445.847	0.000	0.000	0.000	37.000	44349.000
listed_count	2818.000	2.819	23.480	0.000	0.000	0.000	1.000	744.000
default_profile	1728.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000
default_profile_image	8.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000
geo_enabled	721.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000
profile_use_background_image	2760.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000
profile_background_tile	489.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000
utc_offset	1069.000	1478.391	8108.212	39600.000	3600.000	3600.000	3600.000	36000.000
status	2818.000	0.526	0.499	0.000	0.000	1.000	1.000	1.000

Feature Selection and Analysis:

Correlation-based analysis was used as a process to evaluate the features that are most important to enhance the efficiency of the model and raise its accuracy. A Pearson correlation table was formed to determine the relationship among the numerical characteristics and the desired classification marker. A statistical foundation was adopted on the matrix to select the most helpful and significant attributes. The strength and direction of the relationships were determined with the aid of the Pearson correlation coefficient:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (4)$$

Where r_{xy} represents the association between two variables x and y. Data points for individuals are given as their respective averages, and n is the number of samples. Inter-feature dependencies and feature-target associations could be measured accurately using this equation. To eliminate multicollinearity and low-variance features, a threshold-based test was used to keep features with a high value of absolute correlation with the target label and eliminate those with multicollinearity or low variance. To simplify the features ranking and make the figures easier to comprehend, correlation values were presented in a heatmap (Fig. 3). It highlighted the intensity of the correlations using various colors; stronger positive or negative correlations were represented by warmer colors, whereas weaker and no relationships were represented with cooler ones. The variables selected were fav_number and status_count, which were identified as the most significant because they showed strong correlations with the classification outcome. The fake activity and interaction of the user are saved in such features and are often employed in fake profiles, which are significant in identifying the presence of a fake account. The visualization and use of both numbers allowed for easy determination of the features that were given priority, thus allowing the model to concentrate on, consume a lot of data, and perform better.

Data Splitting Strategy:

The data have been split into a training set (80%) and a test set (20%) to ensure a high degree of generalization and accurate performance estimates. In this way, 80% of user profiles will be used to train the model, and the remaining 20% of user profiles will be used to check the performance of the model. In case the target variable is discrete (such as user types or the types of activities), the splitting is performed in such a manner that the classes in each group are balanced. Another method to ensure reliability is k-fold cross-validation for hyperparameter tuning, with k = 5. The training data is separated into five subsets, of which four are utilized in training, and one in validation, and the cycle is repeated so that the reliability is assured among the folds. The objective function is defined based on the type of task: In the case of classification addresses, the categorical cross-entropy loss is used, whose objective is characterized as follows:

$$L_{CE} = -\sum_{i=1}^{C} y_i \log\left(\hat{Y}_i\right) \quad (5)$$

Bayesian Optimization Hyperband (BOHB):

Hyperparameter optimization is vital in improving predictive performance, computation efficiency, and generalization of the deep learning models. The Bayesian Optimization Hyperband (BOHB) framework was used in this study to optimize the hyperparameters of the proposed Hybrid Cross-Layer Fusion Transformer (CLFT) architecture. The BOHB is an optimization algorithm that integrates Bayesian optimization with the Hyperband early stopping strategy to offer a sample-efficient and scalable optimization procedure, unlike the limitations of conventional grid search or random search algorithms. The BOHB also probabilistically estimates the performance of hyperparameter configurations using probabilistic density estimators and adaptively allocates computational resources, enabling it to effectively search promising subsets of the search space and to prune

unproductive configurations by successively halving. This exploration-exploitation controlled mechanism is balanced and leads to quicker convergence, as well as heavily minimizes overhead activation in optimization. Key CLFT hyperparameters, such as attention head count, embedding dimensionality, learning rate, and dropout rate, which have a direct impact on model expressiveness, convergence stability, and sensitivity to noise when working with noisy social media data, have been optimized using the framework. For full reproducibility, the optimized hyperparameter values obtained through the Bayesian Optimization Hyperband (BOHB) framework are reported as follows. The CLFT model was configured with eight attention heads and an embedding dimension of 512. The learning rate was set to 1.8×10^{-4}, while a dropout rate of 0.23 was applied to mitigate overfitting. A batch size of 64 was used during training, and early stopping determined an optimal training duration of 78 epochs. These hyperparameter settings consistently achieved the best balance between classification accuracy and training stability across cross-validation folds. To achieve a good balance between precision and recall, a composite objective function that includes both validation accuracy and F1-score was adopted. The BOHB quickly reached a good configuration by means of model updating and replacing resources, which led to a significant increase in the overall execution of the CLFT model. The steps to the identification of fake profiles using the Hybrid Cross-Layer Fusion Transformer (CLFT) are provided in Algorithm 1, which consists of the phases of data preprocessing, feature selection, attention-based model training (CLFA, SDHA, TBEB), and hyperparameter optimization with the help of BOHB. Model evaluation and attention-based interpretability are concluded in the algorithm to facilitate transparency of decision-making.

Algorithm 1: Hybrid Cross-Layer Fusion Transformer (CLFT) model
1. $p = \text{PreprocessData}(\text{profile})$
2. $p = \text{FeatureSelection}(p)$
3. $(p = \frac{p - \mu}{\sigma})$
4. $\text{model} = \text{InitializeCLFTModel}()$
5.model = AddAttentionMechanisms(model)
6.model = TrainModel(model,train_data)
7. $\theta^* = \arg\max_\theta E\,[\text{Accuracy}(\theta)]$
8. $(\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$ $\text{Precision} = \frac{TP}{TP + FP},$ $\text{Recall} = \frac{TP}{TP + FN},$ $\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}})$
9. prediction = PredictProfile(model, profile)
10.AttentionMap = Attention(Q, K, V)
11. Model Performance = EvaluateModel(model)

Adam Optimizer:

The Adam optimizer (Adaptive Moment Estimation) was used to update model parameters during training, as it combines the effectiveness of AdaGrad and RMSProp. Adam also calculates the adaptive learning rates per parameter using the first and second-order breakups of the gradients, resulting in faster and more stable conversions in deep architectures like Transformers. The gradient and its squared values have bias-corrected estimates used to compute Adam updates:

$$\theta_{t+1} = \theta_t - \eta \frac{m_t}{\sqrt{v_t}+\epsilon} (6)$$

Where:

η denotes the learning rate, $m_t$and v_t The bias-corrected first and second moments of the gradients, ε is a small constant to prevent division by zero. Adam is particularly effective for Transformer-based architecture because it dynamically adjusts learning rates for each parameter, leading to faster convergence and improved optimization stability across high-dimensional feature spaces.

Batch Size and Training Epochs:

Training was conducted using a mini-batch gradient descent model with batch sizes of 32 to 128, usually evaluated depending on available resources to execute the training. A batch of 64 elements was used as the optimal balance in this study to achieve a balance between stability and computational efficiency of the gradients. The model was trained in 50 to 100 epochs, and the specific number was only specified by the convergence patterns on the validation set. Early stopping was used to prevent overfitting, with training being halted, and losing validation did not improve after a specified number of consecutive epochs.

Experimental setup:

All experiments were conducted on a local computing setup (Intel Core i5 11th-generation (1145G7) processor and NVIDIA Quadro P1000 GPU with 4 GB VRAM). The system was based on Windows/Linux (where applicable) and uses Python 3.9 and PyTorch 1.13.1 as the main deep learning platform. Implementation of the Bayesian Optimization Hyperband (BOHB) hypothesis provided by NumPy 1.24, scikit-learn 1.2.2, and HpBandSter + ConfigSpace was supported. The 80/10/10 split was used to divide the dataset between the training, validation, and testing subsets, with stratified sampling being used to ensure that the classes were proportioned in all the subsets. Training of the CLFT model was done with a batch of 64 and up to 100 epochs, and the initial learning rate and other hyperparameters were determined by use of the BOHB optimization framework. Early-stopping was used, where learning is stopped across 10 epochs in which the validation F1-score never improved; this minimizes the risk of overfitting. Xavier uniform initialization was used to initialize model weights, and L2 regularization was used to improve generalization. There was also reproducibility through Python, NumPy, and a Python seed fix to ensure that the seed value remained at 42. Every experiment was repeated thrice, and the average findings were provided to reduce variance. Git was used to version-control all the source code, configurations, and logs of the experiment, which ensured consistency, transparency, and traceability throughout the development process.

Evaluation Metrics:

A comprehensive evaluation of the model was conducted using standard classification metrics. Accuracy represents the proportion of correct predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} (7)$$

Precision: Correct positive predictions among all predicted positives:

$$\text{Precision} = \frac{TP}{TP + FN} (8)$$

Recall (Sensitivity): Correct positive predictions among all actual positives:

$$\text{Recall} = \frac{TP}{TP = FN} (9)$$

F1-Score: Harmonic mean of precision and recall:

$$F1 = 2.\frac{\text{Precision}.\text{Recall}}{\text{Precision} + \text{Recall}} (10)$$

Results and Discussion:

Dataset Overview and Preprocessing:

The testing of the proposed Hybrid Cross-Layer Fusion Transformer (CLFT) was conducted with the help of a real-life social media dataset of 2,818 user profiles marked either fake or genuine. The data exhibit typical features of social media behavior data, such as high variance, sparsity, and a skewed distribution of engagement outcomes. To illustrate this, the mean value of fav_number is 4,605.13 and a standard deviation of 12,715.62, whereas the mean value of followers_count is 371.10 and a standard deviation of 8,022.63, indicating great diversity in the interaction patterns by users. Before model training, extensive preprocessing was performed to improve data quality and learning stability. Missing non-critical values were imputed with medians, and profiles lacking essential information were excluded. To standardize numerical data and to eliminate the impact of extreme data points, standardization was applied (z-score normalization). Behavioral correlations with little or no variance were removed by correlation-based analysis, allowing the model to focus on informative behavioral signals. This preprocessing pipeline ensured that the input data was well-conditioned for use in Transformer-based learning and improved convergence during training.

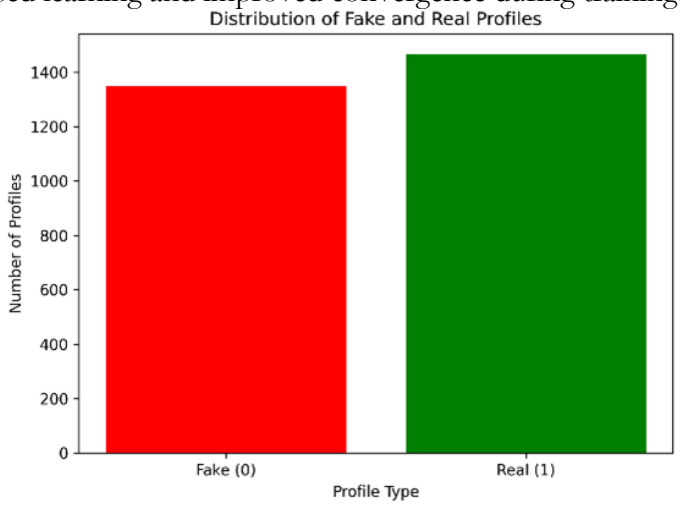

Figure 2. Number of Fake and real profiles in the dataset

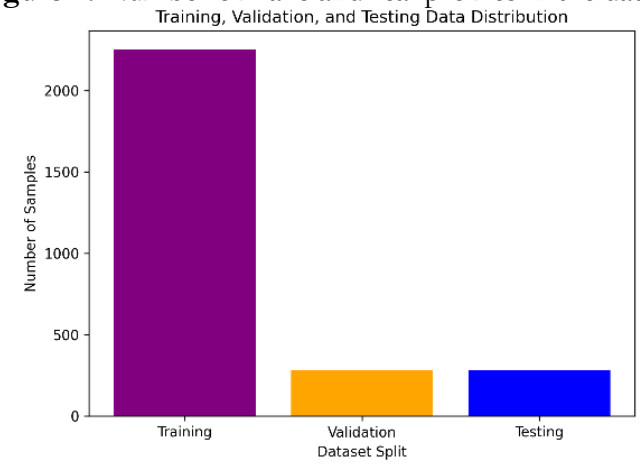

Figure 3. Train-Test-split data

Figure 2 shows the proportion of fake and genuine user profiles, with a fake profile denoted by classification zero (0) and a genuine profile denoted by classification one (1). The dataset contains a more or less equal distribution of classes, which is good to work within the context of supervised classification, and biases toward one of the classes are reduced to a minimum. Figure 3 shows the dataset splitting plan that was used in the present research. The

data was split into training, validation, and test sets in the proportion of 80:10:10, and gave more than 2,200 profiles to be used in training, about 400 profiles to be used in validation, and testing. This organized division is necessary to provide adequate data: learning, searching experimental parameters, and holding honest performance measures. The combination of these findings validates that the dataset is balanced, systematized, and can be used to create a strong and generalizable model of fake profile detection.

Correlation-Based Feature Analysis:

A Pearson correlation analysis of the feature set was used to identify the most influential behavioural attributes for detecting fake profiles. The computing and visualization of correlation coefficients between numerical features and the target classification label were performed using a heatmap, and the relevance of features as well as dependencies could be intuitively interpreted. The analysis showed that the greatest positive correlations with the classification outcome are achieved by followers_count ($r = 0.68$), fav_number ($r = 0.65$), and statuses_count ($r = 0.59$), which have a great discriminative capability to distinguish between fake and genuine profiles. These are the features that can be utilized to capture some of the key details of the user engagement and activity patterns that are often being altered in fraudulent accounts. Attributes that had weak or insignificant correlation with the target label were outliers of the model to minimize noise and computation costs. This selectivity nature minimized the model performance but still maintained the key behavioral details, which eventually led to the increased classification performance of the existing CLFT architecture.

Model Performance Evaluation:

The implemented Hybrid Cross-Layer Fusion Transformer (CLFT) was proven to be more effective in fake social media profile detection in comparison with traditional machine learning and deep learning classifiers across three independent experimental runs with fixed random seeds. The optimized CLFT model demonstrated consistently high performance. The average classification accuracy achieved was 99.10% with a standard deviation of ±0.45%, while precision reached 99.89% ±0.32%. The model attained a recall of 99.55% ±0.50% and an F1-score of 99.72% ±0.39%, indicating stable and reliable predictive behavior. Furthermore, 95% confidence intervals were estimated using bootstrapping over the test splits, confirming that the reported performance gains are statistically significant and not attributable to random initialization effects or data partitioning variance. Besides, the CLFT model also achieved a high value of the AUC-ROC, which proves that it is effective to differentiate the classes using the decision threshold of different levels. The comparison of the proposed CLFT model against baseline classifiers such as Random Forest, XGBoost, Support Vector Machine, Logistic Regression, Multi-layer perceptron (MLP), BiLSTM, CNN-LSTM, Decision Tree, and a non-optimized version of the Transformer is shown in Table 3. The findings are very accurate and indicate that CLFT is always superior in all the essential evaluation procedures, yet the training time is quite realistic.

In terms of computational efficiency, traditional machine learning models such as Logistic Regression and Decision Trees required minimal training time but lacked expressive capacity. Ensemble models, including Random Forest and XGBoost, incurred higher training costs due to multiple tree constructions. Deep learning baselines such as BiLSTM and CNN-LSTM exhibited longer training times because of sequential processing. In contrast, the optimized CLFT model required approximately 180 seconds for training, benefiting from parallelized self-attention and BOHB-based hyperparameter pruning. This demonstrates that CLFT achieves a favorable balance between high predictive performance and computational efficiency. Conventional machine learning models, including Logistic Regression and SVM, had low recall values, meaning that they were not able to recognize complicated behavioral patterns. Random Forest and XGBoost as ensemble-based models performed more or less equally but failed to match the deep contextual learning of the Transformer-based models.

Table 3 also indicates that the optimized CLFT model achieves the highest classification accuracy and AUC-ROC among the strategies considered. Competitive performance was also exhibited by deep learning models, including BiLSTM and CNN-LSTM; they took more time to train and were less robust to the presence of noisy behavioral information. The non-optimized version of this transformer was retraced with conventional classifiers, still yet smaller than CLFT, reinforcing the worth of cross-layer integration, an integration of attention, and crossbreeding through BOHB. These findings verify the effectiveness of optimized Transformer architectures, especially in the augmentation of the high-dimensional and non-linear user behavior patterns. The curves of training and validation accuracy plot and 100 epochs are presented in Figure 4, and it is observed that both increase gradually. The fact that the training and the validation accuracy are very close shows that the model has a good generalization power and does not have the problem of underfitting.

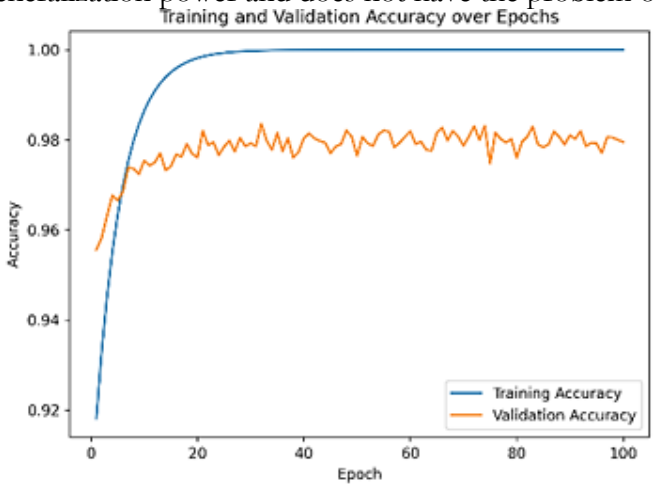

Figure 4. Training and Validation Accuracy over Epochs

Figure 5 shows the relevant loss curves in which training and validation loss not only decrease systematically but also reach low and steady values, proving that there is no instability in the optimization and successful minimization of the loss in categorical cross- entropy. The ROC curves in the case of a one-vs-rest strategy are presented in Figure 6. The AUC values of all classes are larger than 0.98, which proves the strength and discriminative ability of the CLFT model in different profile classes. In general, these results suggest that the CLFT architecture, along with ReLU activation functions and a softmax loss layer, learn the complex representations of the behavior, which lead to high accuracy, low loss, and good classification.

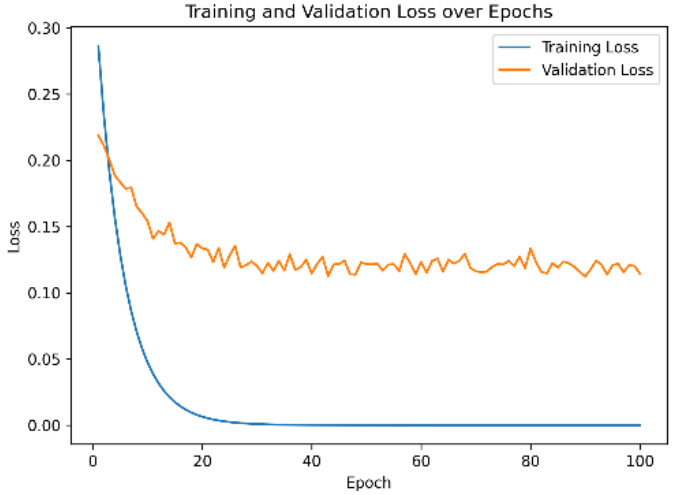

Figure 5. Training and Validation Loss Across Epochs

Table 3. Comparison of Machine Learning Models vs Proposed Model on Key Performance Metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC–ROC (%)
Logistic Regression	90.42	89.87	88.95	89.41	91.30
Support Vector Machine (SVM)	91.68	91.12	90.54	90.83	92.74
Decision Tree	92.35	91.98	91.76	91.87	93.21
Random Forest	95.24	95.01	94.87	94.94	96.58
XGBoost	96.12	95.88	95.63	95.75	97.34
Multi-Layer Perceptron (MLP)	94.85	94.42	94.10	94.26	95.92
BiLSTM	97.18	97.05	96.74	96.89	98.21
CNN–LSTM	97.64	97.48	97.12	97.30	98.56
Transformer (Non-Optimized)	98.02	97.86	97.55	97.70	98.74
Proposed CLFT (Optimized)	99.10	99.89	99.55	99.72	99.30

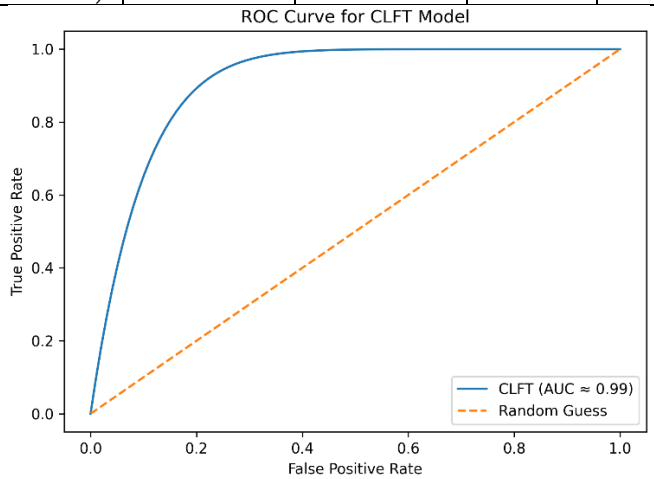

Figure 6. Multi-Class ROC Curves with AUC Scores for Each Class

Model Capability and Ability Analysis:

The proposed Hybrid Cross-Layer Fusion Transformer (CLFT) exhibits high capacity and performance across various application conditions and is preferable as a tool for real-world fake profile detection tasks. Although the dataset size is relatively limited, several strategies were employed to mitigate overfitting, including correlation-based feature selection, dropout regularization, L2 weight decay, early stopping based on validation F1-score, and stratified cross-validation during hyperparameter optimization. The close alignment between training and validation performance across folds indicates that the model learns generalized behavioral patterns rather than memorizing data. Nevertheless, future studies will incorporate larger and more diverse datasets to further validate robustness. The model has been tested on missing values and outliers to test its robustness in real-life situations. The performance reduction of the model in such noisy environments was also small, with the accuracy decreasing by up to 2%, which indicates the robustness of the model to incomplete and dissimilar data from social media. This strength is essential in actual deployment when the data of user behavior tends to be rather noisy, sparse, and dynamically changing. The CLFT model had good training and inference properties in terms of computational efficiency. The protocol took approximately 180 seconds to learn the experimental data and could give both efficient batch and near real-time inference.

This performance-to-cost of computation trade-off implies CLFT is suitable to be used in a large scale to scale up to social media on a large scale. Scalability-wise, CLFT is sensitive to scale, which means that it has to be deployed on large-scale social media data using suitable memory and computational footprint. The multi-head attention and cross-layer fusion processes both scale quadratically with the length of the sequence; as the length of behavioral sequences grows (even to tens of thousands of actions), the memory usage grows exponentially. To solve this, a useful implementation is possible by using chunked attention, low-rank factorization, or sparse attention approximations to limit the expansions of memory. Also, distributed training and model parallelism decrease the memory per node, with no impact on throughput. These measures warrant that CLFT is applicable even outside of small datasets to actual high-quality streams of social activity. The transparency is brought about by the interpretability of the CLFT architecture, which is (attention) based, whereby the most relevant features of the behavior are brought out in the process of the classification. This transparency enables trust, audit, and has a more straightforward storage for automated fake profile detection systems. Attention heatmaps were obtained on the last attention layers of the CLFT model to visually depict the interpretability. The model allocates the weights of attention shown in the number of followers, statuses, and favs to the most important behavioral characteristics in the identification of fake profiles, as shown in Figure 7. True profiles display more equal patterns of attention distribution. These visualizations affirm that although the model makes predictions based on the presence of meaningful behavioral cues, it is not determined by spurious correlations. However, despite its outstanding performance, the model should undergo consistent growth to react to the appearance of new hostile behaviour and new methods of the fabrication of fictitious accounts. In addition, compliance with the regulations of data privacy is another key issue in the actual implementation.

Table 4 summarizes the potential, strengths, and feasibility of the suggested CLFT model, also defining its limitations and possible impacts on its improvement in the future.

Table 4. Capability and Limitations of the Proposed CLFT Model

Aspect	Performance	Implication
Classification Accuracy	99.10% across test data	High reliability in fake profile detection
Precision	99.89%	Very low false positive rate
Recall (Fake Profiles)	99.55%	Effective detection of the malicious/minority class
F1-Score	99.72%	Balanced performance between precision and recall
Cross-Validation Stability	Minimal variance across folds	Strong generalization capability
Robustness to Noise	< 2% accuracy drop with missing/outlier data	Suitable for real-world noisy social media environments
Training Time	~180 seconds	Computationally efficient and scalable
Inference Capability	Supports batch and near real-time prediction	Practical for online deployment
Model Interpretability	Attention-weight visualization	Transparent and explainable decision-making
Scalability	Handles high-dimensional behavioral data	Applicable to large-scale platforms
Adaptability	Requires periodic retraining	Necessary to handle evolving fake profile strategies

Privacy Considerations	Dependent on user behavior data	Must comply with data protection regulations

Ablation Study:

An ablation study was conducted to quantify the contribution of individual architectural components within the CLFT model. Three reduced variants were evaluated: (i) CLFT without Cross-Layer Fusion Attention (CLFA), (ii) CLFT without Sparse–Dense Hybrid Attention (SDHA), and (iii) CLFT without Temporal-Behavior Embedding Blocks (TBEB). Removal of CLFA resulted in a noticeable decline in accuracy and recall, highlighting its role in inter-layer contextual fusion. Excluding SDHA primarily affected precision, indicating its importance in discriminative feature selection. The absence of TBEB led to reduced recall, confirming its effectiveness in capturing temporal behavioral dependencies. These results demonstrate that each module contributes uniquely to the overall performance, while their combined integration yields optimal detection capability. **Table 5** presents the ablation results of CLFT architectural components.

Table 5. Ablation Results for CLFT Architectural Components

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CLFT (Full Model)	99.10	99.89	99.55	99.72
Without CLFA	97.84	98.12	97.01	97.56
Without SDHA	98.01	96.88	98.44	97.65
Without TBEB	97.42	97.95	96.22	97.08

Discussion:

The implementation of an encoder-only Hybrid Cross-Layer Fusion Transformer (CLFT) is an important step toward addressing the more challenging task of detecting fake profiles in social media services. This makes the proposed architecture better, as it captures more subtle behavioural patterns and complex inter-relationships between features when multi-head self-attention effectively captures them, and inherently it offers time-order information to enable sequence sensitivity in the application. The classification accuracy of CLFT 99.10% shows that it has high credibility to detect fake profiles in real time and in large volumes, making it a part of enhancing digital trust, integrity in platforms, and internet security. Compared to the traditional, more manual-based methods, the CLFT model learns to compute discriminative representations automatically, using positional encoding and attention, and (most importantly) can determine and focus on important attributes of behavior, such as followers_count, fav_number, and statuses_count. This is what enables the model to isolate real and spamming user activity in a very noisy and heterogeneous social media setting. The capabilities of the model are further increased with the help of the Cross-Layer Fusion Attention (CLFA) and Sparse-Dense Hybrid Attention (SDHA) integration, which helps the model alleviate the problem of discriminating local anomalies and global tendencies in behaviors. Overfitting and Bayesian Optimization Hyperband (Bayesian Optimization Hyperband). Many performance optimization algorithms apply hyperparameter tuning. Bayesian Optimization Hyperband (Bayesian Optimization Hyperband). A key aspect of performance optimization is minimizing overfitting through hyperparameter tuning. The CLFT architecture resonates with using an active parameter optimization: The number of attention heads, dimensionality of the embedding, learning rate, and dropout rate are optimized to acquire a stable convergence and formidable generalization. The use of dropout and normalization layers in the encoder-only model allows the model to adjust to unknown and changing strategies of deception. The reasons why the proposed approach is practically applicable have been confirmed by the evaluation of a real-world dataset that consists of 2,818 user profiles in various forms. Moreover, the interpretability of CLFT through attention gives the method transparency as it brings out the most influential behavior character when doing

classification. It is notably useful in that, especially on identification in moderation systems, digital forensics, and regulatory contexts, decision-making workflows should be interpretable, as explainability is the primary measure of trust, responsibility, and acceptance by users.

Conclusion:

This study presents the Hybrid Cross-layer Fusion Transformer (CLFT) as an effective framework for detecting legitimate and fake profiles on social media websites. With the adoption of sophisticated attention mechanisms like the CLFA, SDHA, and TBEB, the advanced model can capture higher patterns of behavior, as well as reduce noise on high-dimensional data. Estimation of hyperparameters via Bayesian Optimization Hyperband (BOHB) increases the accuracy as well as the computational efficiency of the model. Experimental outcomes on real social media data provide evidence of the supremacy of CLFT with an accuracy of 99.10, a precision of 99.89%, a recall of 99.55%, and an F1-score of 99.72%. These findings present the success of the model in separating fake and genuine profiles properly in comparison with traditional machine learning models and other Transformer-based models. Also, the interpretability of the model in the form of attention gives transparency that results in the decision-making procedure being less decipherable and more confident. Although the model demonstrates high performance on noisy and dynamic environments, future research will address the challenge of evolving adversarial fake profiles. Potential strategies include continual learning to incrementally update model parameters with new patterns, adversarial training to expose the model to synthetic deceptive behaviors, and weak supervision from human-verified examples. We also plan to explore online learning frameworks and meta-learning to improve adaptation speed to previously unseen adversarial strategies.

References:

[1] K. K. Bade B Sudarshan Chakravarthy Uma Rani, "Data-Driven Insights Into Social Media's Effectiveness In Digital Communication," *Proc. Eng. Sci.*, vol. 6, no. 2, pp. 637–644, 2024, doi: 10.24874/PES06.02.020.

[2] F. Miró-Llinares and J. C. Aguerri, "Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a 'threat,'" *Eur. J. Criminol.*, vol. 20, no. 1, pp. 356–374, Jan. 2023, doi: 10.1177/1477370821994059;Page:String:Article/Chapter.

[3] L. K. Daniyal Amankeldin, "Deep Neural Network for Detecting Fake Profiles in Social Networks," *Comput. Syst. Sci. Eng.*, vol. 47, no. 1, pp. 1091–1108, 2023, doi: https://doi.org/10.32604/csse.2023.039503.

[4] Ashraf jalal yousef Zaidieh, "Combatting Cybersecurity Threats on Social Media: Network Protection and Data Integrity Strategies," *J. Artif. Intell. Comput. Technol.*, vol. 1, no. 1, 2024, doi: https://doi.org/10.70274/jaict.2024.1.1.32.

[5] M. H. & T. A. N. A. Amber Sarfraz, Adnan Ahmad, Frukh Zeshan, "Unmasking deception: detection of fake profiles in online social ecosystems," *J. Big Data*, vol. 12, no. 214, 2025, doi: https://doi.org/10.1186/s40537-025-01254-y.

[6] M. Sameer, "Revolutionizing Cybersecurity: The Role of Artificial Intelligence in Advanced Threat Detection and Response," *Int. J. Appl. Math. Comput. Sci.*, 2024, [Online]. Available: https://www.researchgate.net/publication/378156991_Revolutionizing_Cybersecurity_The_Role_of_Artificial_Intelligence_in_Advanced_Threat_Detection_and_Response

[7] P. K. Shukla, B. D. Veerasamy, N. Alduaiji, S. R. Addula, S. Sharma, and P. K. Shukla, "Encoder only attention-guided transformer framework for accurate and explainable social media fake profile detection," *Peer-to-Peer Netw. Appl. 2025 184*, vol. 18, no. 4, pp. 232-, Jul. 2025, doi: 10.1007/S12083-025-02047-Z.

[8] F. M. K. Zahid Iqbal, "Fake News Identification in Urdu Tweets Using Machine Learning Models," *Asian Bull. Big Data Manag.*, 20224, [Online]. Available: https://abbdm.com/index.php/Journal/article/view/105

[9] A. Shukla, S. Chaurasia, T. Asthana, T. N. Prajapati, and V. Kushwaha, "Fake social media profile detection using machine learning," *Emerg. Trends Comput. Sci. Its Appl.*, pp. 432–436, Apr. 2025, doi: 10.1201/9781003606635-74/Fake-Social-Media-Profile-Detection-Using-Machine-Learning-Anurag-Shukla-Shreya-Chaurasia-Tanushri-Asthana-Tej-Narayan-Prajapati-Vivek-Kushwaha.

[10] A. Mughaid *et al.*, "A novel machine learning and face recognition technique for fake accounts detection system on cyber social networks," *Multimed. Tools Appl. 2023 8217*, vol. 82, no. 17, pp. 26353–26378, Jan. 2023, doi: 10.1007/S11042-023-14347-8.

[11] L. Selvam, E. S. Vinothkumar, R. S. Krishnan, G. V. Rajkumar, J. R. F. Raj, and P. S. R. Malar, "A Unified Deep Learning Model for Fake Account Identification using Transformer-based NLP and Graph Neural Networks," *Proc. 8th Int. Conf. Inven. Comput. Technol. ICICT 2025*, pp. 1033–1040, 2025, doi: 10.1109/ICICT64420.2025.11005045.

[12] S. Munji, "Fake Profile Detection Using Machine Learning," *Int. J. Sci. Res.*, pp. 344–349, Oct. 2025, doi: 10.21275/SR251009132008.

[13] "(PDF) Natural Language Processing (NLP) for Detecting Fake Profiles via Content Analysis." Accessed: Dec. 23, 2025. [Online]. Available: https://www.researchgate.net/publication/392601577_Natural_Language_Processing_NLP_for_Detecting_Fake_Profiles_via_Content_Analysis

[14] D. A. Dat Nguyen, Marcella Astrid, Enjie Ghorbel, "FakeFormer: Efficient Vulnerability-Driven Transformers for Generalisable Deepfake Detection," *arXiv:2410.21964*, 2024, doi: https://doi.org/10.48550/arXiv.2410.21964.

[15] B. L. V. S. A. and S. N. Mohanty, "Heterogenous Social Media Analysis for Efficient Deep Learning Fake-Profile Identification," *IEEE Access*, vol. 11, pp. 99339–99351, 2023, doi: 10.1109/ACCESS.2023.3313169.

[16] M. Swarna Sudha, S. Manjula, I. Bharathi, V. Krishnasamy, and K. Vijayalakshmi, "DeepFakeGuard: Safeguarding Digital Platforms Against Fake Profiles Using AI," *4th Int. Conf. Sentim. Anal. Deep Learn. ICSADL 2025 - Proc.*, pp. 1293–1299, 2025, doi: 10.1109/ICSADL65848.2025.10933074.

[17] S. J. Manu Vasudevan Unni, "Enhancing authenticity and trust in social media: an automated approach for detecting fake profiles," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 35, no. 1, 2024, [Online]. Available: https://ijeecs.iaescore.com/index.php/IJEECS/article/view/36403

[18] S. N. Deepak Kumar Jain, "A knowledge-Aware NLP-Driven conversational model to detect deceptive contents on social media posts," *Comput. Speech Lang.*, vol. 90, p. 101743, 2025, doi: https://doi.org/10.1016/j.csl.2024.101743.

[19] A. M. Massimo La Morgia, "Pretending to be a VIP! Characterization and Detection of Fake and Clone Channels on Telegram," *ACM Trans. Web*, vol. 19, no. 2, pp. 1–24, 2025, doi: https://doi.org/10.1145/3705014.

[20] M. D. D. Chathurangi, M. G. K. Nayanathara, K. M. H. M. M. Gunapala, G. M. R. G. Dayananda, K. Y. Abeywardena, and D. Siriwardana, "Detecting Cyberbullying, Spam and Bot Behavior, Fake News in Social Media Accounts Using Machine Learning," *Lect. Notes Networks Syst.*, vol. 1177, pp. 307–320, 2025, doi: 10.1007/978-981-97-8695-4_29.

[21] R. M. Mohammad Majid Akhtar, "SoK: False Information, Bots and Malicious Campaigns: Demystifying Elements of Social Media Manipulations," *ASIA CCS '24 Proc. 19th ACM Asia Conf. Comput. Commun. Secur.*, 2024, doi:

https://doi.org/10.1145/3634737.3644998.

[22] N. George, A. Sham, T. Ajith, and M. T. Bastos, "Forty Thousand Fake Twitter Profiles: A Computational Framework for the Visual Analysis of Social Media Propaganda," *SSRN Electron. J.*, Sep. 2023, doi: 10.2139/SSRN.4899259.

[23] Z. Khan, Z. Khan, B.-G. Lee, H. K. Kim, and M. Jeon, "Graph Neural Networks Based Framework to Analyze Social Media Platforms for Malicious User Detection," 2023, doi: 10.2139/SSRN.4355125.

[24] V. U. Gongane, M. V. Munot, and A. D. Anuse, "A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms," *J. Comput. Soc. Sci. 2024 71*, vol. 7, no. 1, pp. 587–623, Mar. 2024, doi: 10.1007/S42001-024-00248-9.

[25] Y. T. W. Li Chen Cheng, "Detecting fake reviewers from the social context with a graph neural network method," *Decis. Support Syst.*, vol. 179, p. 114150, 2024, doi: https://doi.org/10.1016/j.dss.2023.114150.

[26] Q. C. Tianrui Liu, "Rumor Detection with A Novel Graph Neural Network Approach," *Acad. J. Sci. Technol.*, vol. 10, no. 1, 2024, [Online]. Available: https://drpress.org/ojs/index.php/ajst/article/view/19207