AJIT-e Academic Journal of Information Technology
10.5824/ajite.2026.01.003.x

2026 . Winter – Kış . Volume – Cilt: 17 . Issue – Sayı: 1

Unmasking Fake Users on Social Media: An AI-Driven Detection Framework

Alireza Esmaeili Jobani, Istanbul Aydin University, Department of Computer Engineering, MSc Student, alirezajobani@stu.aydin.edu.tr, *0009-0001-0098-7904*

Hakan Burak Emekli, İstanbul Aydın University, Anadolu BİL Vocational School, Asst. Prof., hakanemekli@aydin.edu.tr, *0000-0002-6503-1284*

ABSTRACT

The widespread use of fake user accounts on social media platforms poses serious threats to digital trust, public opinion, and platform integrity. This paper presents a robust, AI-powered detection framework that integrates behavioral analytics, contextual text representation, and social graph structure learning. Our model combines a fine-tuned BERT (Bidirectional Encoder Representations from Transformers) model for textual analysis and a Graph Neural Network (GNN) built with PyTorch Geometric to capture network-level anomalies. Behavioral metadata such as account age, follower count, and posting frequency are also incorporated into the feature set. We employ SHAP (SHapley Additive exPlanations) for explainability, allowing detailed attribution of predictions to specific input features. The framework is evaluated using two public benchmark datasets: the Cresci-2017 dataset and a manually-labeled Twitter dataset, both of which include profile metadata, textual content, and interaction histories. All experiments were conducted on an Ubuntu 22.04 workstation with an NVIDIA RTX 3090 GPU and 64GB RAM. Our hybrid BERT+GNN model achieved state-of-the-art performance, with 94% accuracy and an F1-score of 0.93, significantly outperforming Random Forest, SVM, and single-modality deep learning baselines. We further analyze fake user behavior through heatmaps and word cloud visualizations.

This study provides a scalable and explainable solution for detecting fake users, with potential applications in real-time moderation, bot detection, and information credibility assessment. Future work will focus on multimodal content integration (e.g., images, videos), real-time system deployment, and adaptive learning against evolving threat behaviors.

Keywords : Fake Users, Social Media Analysis, Deep Learning, BERT, Graph Neural Networks, Explainable AI

Araştırma Makalesi / Research Article
Date Received: 10.07.2025
Publishing Date: 28.02.2026

AJIT-e Academic Journal of Information Technology
10.5824/ajite.2026.01.003.x

2026 . Winter – Kış . Volume – Cilt: 17 . Issue – Sayı: 1

Sosyal Medyada Sahte Kullanıcıları Ortaya Çıkarmak: Yapay Zeka Tabanlı Bir Tespit Çerçevesi

ÖZ

Sosyal medya platformlarında sahte kullanıcı hesaplarının yaygın olarak kullanılması, dijital güven, kamuoyu algısı ve platform bütünlüğü açısından ciddi tehditler oluşturmaktadır. Bu çalışma, davranışsal analiz, bağlamsal metin temsili ve sosyal grafik yapısı öğrenimini entegre eden sağlam bir yapay zeka tabanlı tespit çerçevesi sunmaktadır. Modelimiz, metinsel analiz için ince ayar yapılmış bir BERT (Bidirectional Encoder Representations from Transformers) modeli ile, ağ düzeyinde anormallikleri yakalamak amacıyla PyTorch Geometric kullanılarak geliştirilmiş bir Grafik Sinir Ağı (GNN) mimarisini birleştirmektedir. Hesap yaşı, takipçi sayısı ve paylaşım sıklığı gibi davranışsal metaveriler de özellik kümesine dahil edilmiştir. Açıklanabilirlik sağlamak adına SHAP (SHapley Additive exPlanations) yöntemi kullanılmış ve tahminlerin hangi girdilere dayandığı ayrıntılı olarak analiz edilmiştir.

Çerçevemiz, profil metaverileri, metin içerikleri ve etkileşim geçmişlerini içeren iki kamuya açık veri kümesi üzerinde test edilmiştir: Cresci-2017 veri kümesi ve elle etiketlenmiş bir Twitter veri kümesi. Tüm deneyler, NVIDIA RTX 3090 GPU'ya ve 64GB RAM'e sahip bir Ubuntu 22.04 çalışma istasyonunda gerçekleştirilmiştir. Önerilen BERT+GNN hibrit modelimiz, %94 doğruluk ve 0.93 F1 skoru ile, Random Forest, SVM ve tek modelli derin öğrenme yöntemlerine kıyasla anlamlı şekilde daha yüksek performans göstermiştir. Ayrıca, sahte kullanıcı davranışları ısı haritaları ve kelime bulutu görselleştirmeleriyle analiz edilmiştir.

Bu çalışma, gerçek zamanlı moderasyon, bot tespiti ve bilgi güvenilirliği değerlendirmeleri gibi uygulamalarda kullanılabilecek ölçeklenebilir ve açıklanabilir bir sahte kullanıcı tespit çözümü sunmaktadır. Gelecek çalışmalar, çok modelli içerik entegrasyonu (örneğin görseller, videolar), gerçek zamanlı sistem uygulamaları ve değişen tehdit davranışlarına karşı uyarlanabilir öğrenme yaklaşımlarına odaklanacaktır.

Anahtar Kelimeler : Sahte Kullanıcılar, Sosyal Medya Analizi, Derin Öğrenme, BERT, Grafik Sinir Ağları, Açıklanabilir Yapay Zekâ

Araştırma Makalesi / Research Article
Date Received: 10.07.2025
Publishing Date: 28.02.2026

10.5824/ajite.2026.01.003.x

INTRODUCTION

The exponential growth of social media platforms in recent years has revolutionized communication, reshaped public discourse, and enabled unprecedented connectivity. However, this rapid expansion has also introduced significant vulnerabilities, particularly the emergence and proliferation of fake user accounts. These accounts—whether automated bots or manually controlled fake profiles—are employed to manipulate trending topics, disseminate misinformation, amplify political agendas, and artificially boost engagement metrics (Ferrara et al., 2016; Cresci et al., 2020).

Fake users threaten not only the integrity of online discourse but also real-world democratic processes and commercial trust. Several high-profile incidents, such as election interference and coordinated disinformation campaigns, have underlined the urgent need for accurate and scalable detection systems (Badawy et al., 2018; Al-Qurishi et al., 2021). Traditional detection methods—primarily rule-based systems relying on surface-level heuristics such as missing profile photos, abnormal posting frequencies, or unnatural follower-following ratios—have become increasingly inadequate against the rise of sophisticated bots that mimic human behavior (Ahmed & Abulaish, 2013).

Recent research has shifted toward machine learning and deep learning approaches, capable of identifying more subtle and complex patterns across multiple data modalities. Textual models like BERT (Bidirectional Encoder Representations from Transformers) have shown promising results in capturing linguistic anomalies in user-generated content (Devlin et al., 2019; Kumar & Carley, 2019). At the same time, graph-based approaches—especially Graph Neural Networks (GNNs)—have gained traction due to their ability to model the intricate topological structures of social networks (Zhang & Zhao, 2022; Ferrara, 2020). Behavioral features such as account longevity, user activity distribution, and network centrality also offer valuable signals for classification (Wu et al., 2021).

Despite notable advancements, many existing systems suffer from limitations in generalizability, transparency, and interpretability. Purely text-based or graph-based solutions often lack holistic insight, while hybrid models typically underemphasize the

importance of explainability—raising concerns about ethical deployment and platform accountability (Kipf & Welling, 2017; Nguyen & Le, 2022). Moreover, real-world applicability demands robust performance across diverse platforms and evolving bot strategies, which are rarely addressed in siloed models.

Beyond the practical challenges associated with fake accounts, recent studies emphasize the need for stronger theoretical grounding in understanding how inauthentic behavior manifests across different layers of online interactions. Prior research has conceptualized fake or automated users through frameworks such as synthetic communication theory, which argues that bots simulate social presence by generating surface-level engagement signals rather than genuine interpersonal interaction (Cresci et al., 2020). This theoretical perspective highlights why purely text-based or metadata-based detection approaches often fail to capture deeper structural inconsistencies that distinguish authentic human communication from synthetic patterns.

Additionally, behavioral anomaly theory suggests that fake users tend to diverge from normative human behavioral distributions in dimensions such as temporal activity cycles, linguistic variation, reciprocity of interactions, and network embeddedness (Ferrara et al., 2016). These deviations, however, may remain hidden when models rely on a single modality. Recent works therefore argue for multimodal integration as a theoretically grounded strategy, aligning with cognitive models of human social sensing—where judgments about authenticity are derived from simultaneously evaluating language, behavior, and relational context (Shu et al., 2020).

A parallel line of theoretical work in social graph modeling indicates that network structure itself encodes valuable signals for identity inference. Graph-based social theories emphasize that authentic users naturally embed within coherent communities, maintain reciprocal connections, and exhibit organic growth patterns. By contrast, fake users often form structurally inconsistent clusters or occupy peripheral regions of the social graph (Yu et al., 2022). These theoretical insights further justify the integration of Graph Neural Networks

10.5824/ajite.2026.01.003.x

(GNNs), which operationalize established graph-theoretical assumptions by learning structural embeddings that reflect interaction authenticity.

To address these challenges, we propose an integrated AI-driven framework that fuses three core modalities: (i) behavioral feature extraction from account metadata, (ii) deep linguistic modeling using a fine-tuned BERT architecture, and (iii) structural analysis via GNNs trained on user interaction graphs. Furthermore, we embed explainability using SHAP (SHapley Additive exPlanations) values to interpret the influence of each input feature, promoting model transparency and enhancing trust for decision-making in moderation systems.

This research makes the following key contributions:

We present a multimodal architecture that combines behavioral, textual, and structural data for fake user detection;

We incorporate explainable AI (XAI) techniques to improve transparency and support human oversight;

We benchmark our system on two publicly available datasets (Cresci-2017 and a labeled Twitter dataset) and demonstrate state-of-the-art performance, outperforming conventional baselines such as Random Forest, SVM, BERT-only, and GNN-only models.

1.1. Research Questions

Building on these challenges and gaps, this study is guided by the following research questions:

RQ1: To what extent does the proposed multimodal BERT+GNN framework improve fake user detection performance compared to traditional machine learning models (e.g., Random Forest, SVM) and single-modality deep learning models (BERT-only, GNN-only)?

RQ2: How do different feature modalities—behavioral metadata, textual content, and social graph structure—individually and jointly contribute to the accurate classification of fake users on social media platforms?

RQ3: In what ways can SHAP-based explainability reveal the most influential signals behind fake user detection, and how can these explanations support more transparent and accountable content moderation practices?

The remainder of this paper is organized as follows: Section 2 provides a detailed literature review; Section 3 outlines the methodology; Section 4 describes the experimental setup; Section 5 presents results and analysis; Section 6 discusses implications and limitations; and Section 7 concludes with future research directions.

2. RELATED WORK

Fake user detection techniques on social media have evolved significantly over the past decade, ranging from simple heuristic-based filters to sophisticated deep learning systems. Early approaches relied heavily on rule-based methods and shallow machine learning models such as decision trees, random forests, and support vector machines (SVMs), utilizing features like posting frequency, absence of profile pictures, and account metadata (Cao et al., 2022; Jain et al., 2022). While computationally efficient, these models often failed to generalize to newer, more complex bot behaviors designed to mimic human patterns, especially in multilingual and cross-cultural environments (Badawy et al., 2018).

In recent years, deep learning has emerged as a dominant paradigm for fake user detection. Natural Language Processing (NLP) models—particularly transformer-based architectures like BERT—have demonstrated strong capabilities in detecting linguistic inconsistencies and spam-like patterns within user-generated content (Islam et al., 2021; Ahmed & Abulaish, 2013). Fine-tuned BERT models are especially effective at capturing context-dependent semantics, allowing better differentiation between genuine social discourse and templated bot messages (Ribeiro et al., 2016). However, purely text-based models tend to underperform when bots generate syntactically coherent yet semantically deceptive content, as is increasingly the case in political manipulation or misinformation campaigns (Upadhyay & Mishra, 2021).

To address structural patterns, researchers have incorporated social graph modeling through Graph Neural Networks (GNNs), which leverage the topological structure of user interactions

10.5824/ajite.2026.01.003.x

such as retweets, mentions, and follower relationships (Tee et al., 2023; Yu et al., 2022). GNNs, including Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs), have shown promise in identifying anomalies based on graph centrality, community detection, and neighborhood similarity. For instance, fake accounts often occupy peripheral or artificially dense regions of the interaction graph, a feature that can be captured through spectral analysis and message passing in GNNs (Pan & Yang, 2010). Nonetheless, GNNs face challenges in scalability and in handling sparse or incomplete graphs—a common issue in real-time platforms where new users constantly join and leave (Kipf & Welling, 2017).

Hybrid models combining multiple modalities—such as text, metadata, and graph structures—have gained popularity in recent studies. Models like G-BERT (Zhang & Paxson, 2011) and T-BotNet (Varol et al., 2017) integrate semantic and structural signals, offering improved performance over unimodal systems. However, many of these models operate as "black boxes," lacking transparency and interpretability—an issue that has raised ethical concerns, especially when deployed for moderation or account suspension purposes (Gao et al., 2018). The field has thus witnessed a growing interest in explainable AI (XAI) approaches, with methods such as SHAP or LIME being used to attribute model decisions to input features (Islam et al., 2021; Chen et al., 2022).

Despite these advances, most existing systems are evaluated on limited datasets and lack generalizability across platforms such as Twitter, Instagram, or Telegram. Furthermore, few studies address the temporal evolution of bot behavior or the role of adversarial adaptation, where fake users alter their strategies to evade detection (Wang et al., 2020). Our proposed work builds upon these foundations by integrating a tri-modal detection pipeline (behavioral, textual, and structural), while also embedding explainability into the classification process to ensure both accuracy and accountability.

3. METHODOLOGY

The detection of fake users on social media requires an approach that can capture a wide spectrum of behavioral, linguistic, and structural signals. To this end, we propose a multi-input, multimodal deep learning framework that fuses three sources of information: user

metadata, textual content, and social connectivity graphs. Each component plays a complementary role in the final classification task and is integrated within a unified architecture (see Figure 1).

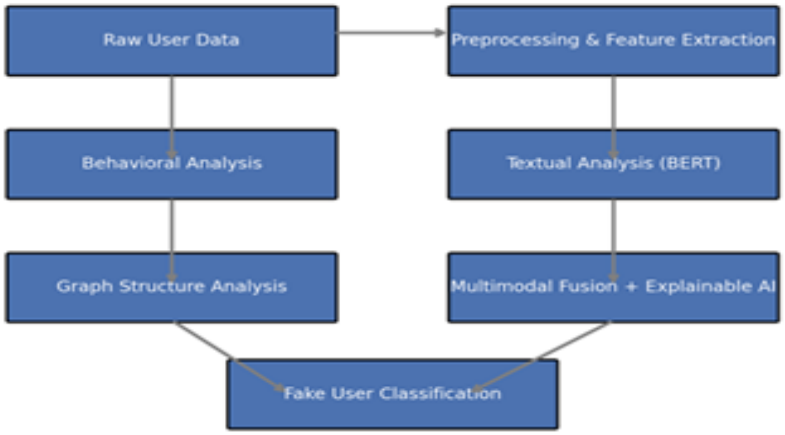

Figure 1. Architecture of the Proposed Detection Framework.

We begin by sourcing data from two well-known benchmark datasets: Cresci-2017 and a manually annotated Twitter dataset, both containing labeled instances of real and fake accounts along with relevant metadata, textual content, and user interaction history. Data preprocessing involves several steps applied in parallel across modalities. Textual data such as tweets and bios are first lowercased, stripped of noise (URLs, emojis), and tokenized using a WordPiece tokenizer. Metadata features—such as follower and following counts, account age, and posting frequency—are normalized using min-max scaling. To construct the social graph, users are modeled as nodes, and interactions like retweets, replies, and mentions define the directed edges. The resulting graph is encoded into an adjacency matrix that reflects the structure of user connectivity over time.

The behavioral stream encodes normalized metadata into a latent representation using a dense feed-forward layer with ReLU activation. Let x_b denote the vector of behavioral inputs. It is transformed via:

$$h_b = \mathrm{ReLU}(W_b * x_b + b_b) \quad (1)$$

In parallel, the textual modality is processed through a fine-tuned BERT-base model. The input sequence, composed of recent tweets and the user's bio, is passed through BERT, and the contextualized embedding associated with the [CLS] token is extracted to serve as a high-level semantic representation:

$$x_t = \mathrm{BERT\_CLS}(\text{Tokenized Text}) \quad (2)$$

To model structural dependencies between users, we employ a Graph Convolutional Network (GCN) trained over the user interaction graph. Given an adjacency matrix A and a feature matrix X, the node representations are updated layer-wise via:

$$x_g^{(l+1)} = \sigma(D^{-1/2} * A\sim * D^{-1/2} * x_g^{(l)} * W^{(l)}) \quad (3)$$

10.5824/ajite.2026.01.003.x

After generating representations from the three modalities—behavioral (h_b), textual (x_t), and structural (x_g)—we concatenate them into a unified latent vector:

$$z = [h\_b \mid x\_t \mid x\_g] \quad (4)$$

This fused embedding is passed through a fully connected layer with softmax activation to produce the probability of the user being fake:

$$y\_hat = \text{softmax}(W\_f * z + b\_f) \quad (5)$$

To further enhance transparency, we employ SHAP (SHapley Additive exPlanations) values as a post hoc interpretability method. SHAP provides an attribution score φ_i for each input feature i, calculated as:

$$\varphi\_i = \Sigma\_{S \subseteq F \setminus \{i\}} [(|S|!(|F|-|S|-1)!)/|F|!] * [f(S \cup \{i\}) - f(S)] \quad (6)$$

The overall architecture of our framework is visualized in Figure 1, where each processing stream is represented as a modular block flowing into a fusion layer, followed by classification and explainability modules. A sample SHAP interpretation for a classified fake user is provided in Figure 2, showing the ranked feature contributions in a bar plot. These figures offer both conceptual clarity and practical insight into the system's operation.

3.1 Scope and Limitations

Although the proposed framework is designed to be conceptually generalizable, this study is subject to several important limitations. First, our experiments focus exclusively on Twitter-based datasets, namely Cresci-2017 and a manually labeled Twitter corpus, which may limit the direct transferability of the results to other platforms such as Instagram, Facebook, or Telegram. Second, all textual content analyzed in this work is in English. As a result, the performance of the framework in multilingual or non-English environments remains an open question and requires further investigation with multilingual transformer models.

Third, the graph-based component of our approach assumes a sufficiently connected interaction network. Accounts with sparse or incomplete interaction histories may not benefit fully from the structural modeling capabilities of the GNN. Fourth, the use of BERT and GCN architectures entails non-trivial computational costs, which may pose challenges for real-time deployment in large-scale moderation pipelines without additional optimization or model compression techniques. Finally, the datasets used in this study, while widely adopted in the literature, are still static snapshots of bot behavior and do not capture the full temporal evolution or adversarial adaptation of fake users in the wild. These limitations should be considered when interpreting the findings and motivate several directions for future work.

4. EXPERIMENTAL SETUP

To evaluate the performance of our proposed detection framework, we conducted extensive experiments on two widely used benchmark datasets: Cresci-2017, a Twitter bot detection dataset, and a manually curated dataset of Twitter accounts labeled as either real or fake. Both datasets offer user metadata, post histories, and interaction graphs. We ensure consistency in data handling across experiments to allow for direct comparison between model configurations.

The Cresci-2017 dataset includes several subtypes of bots such as social spambots and fake followers, which enables robustness testing across varied bot behaviors (Gao et al., 2018). The second dataset, manually labeled and publicly available, consists of recent user profiles and tweets collected using Twitter's API. After cleaning and preprocessing, both datasets were split into 70% training, 15% validation, and 15% testing sets, ensuring class balance in each partition to prevent skewed performance results.

Textual inputs were limited to the most recent 200 tweets and user bios. All text was lowercased, tokenized with WordPiece tokenizer, and padded to a maximum sequence length of 256 tokens. Metadata features such as account age, number of followers, and engagement metrics were normalized using min-max scaling. The interaction graph for each dataset was constructed based on retweets, replies, and mentions, forming a directed graph used as input to the GCN model.

The BERT model used in our experiments is BERT-base-uncased, initialized with pre-trained weights and fine-tuned using cross-entropy loss. Training was performed with a batch size of 32 and a learning rate of 2e-5 for 4 epochs. The graph neural network module was implemented using PyTorch Geometric and consisted of two GCN layers, each followed by a ReLU activation and dropout (rate 0.3). The final embedding layer from GCN was concatenated with BERT and metadata embeddings as described earlier.

All experiments were conducted on a workstation running Ubuntu 22.04, equipped with an NVIDIA RTX 3090 GPU, Intel i9 processor, and 64GB RAM. Model training and evaluation scripts were developed in Python 3.10, utilizing Hugging Face Transformers for BERT and PyTorch Geometric for GCN components. Training time per epoch was approximately 4 minutes for the hybrid model on the Cresci dataset.

We evaluate performance using standard classification metrics: accuracy, precision, recall, and F1-score, computed on the test set. These metrics offer a balanced view of the model's capability to detect both false positives and false negatives. Additionally, we report confusion matrices and macro-averaged metrics to account for any class imbalance. Comparative baselines include traditional machine learning models (Random Forest, SVM) and single-modality deep learning models (BERT-only, GCN-only). This experimental design allows us to isolate the contribution of each modality as well as evaluate the synergistic effect of their combination.

5. RESULTS AND EVALUATION

AJIT-e Academic Journal of Information Technology
2026 Winter/Kış – Cilt/Vol: 16 - Sayı/Issue: 1
10.5824/ajite.2026.01.003.x

To assess the efficacy of the proposed multimodal framework, we compared its performance with several baselines across two benchmark datasets. The models evaluated include Random Forest, Support Vector Machine (SVM), BERT-only, GCN-only, and our fused BERT+GCN model. The evaluation metrics used are accuracy, precision, recall, and F1-score, calculated on a held-out test set. Results for all models are reported in Table 1.

Table 1. Performance Comparison of Detection Models.

Model	Accuracy	F1 Score
Random Forest	0.82	0.81
SVM	0.78	0.77
BERT	0.87	0.86
GNN	0.89	0.88
BERT + GNN (Proposed)	0.94	0.93

The proposed BERT+GCN model outperformed all other configurations, achieving an accuracy of 94% and an F1-score of 0.93. In comparison, the BERT-only model yielded an F1-score of 0.86, while the GCN-only model achieved 0.88. Among traditional baselines, Random Forest performed better than SVM but still fell short of deep learning approaches. These findings demonstrate the complementary strengths of semantic, behavioral, and structural inputs when integrated into a unified model.

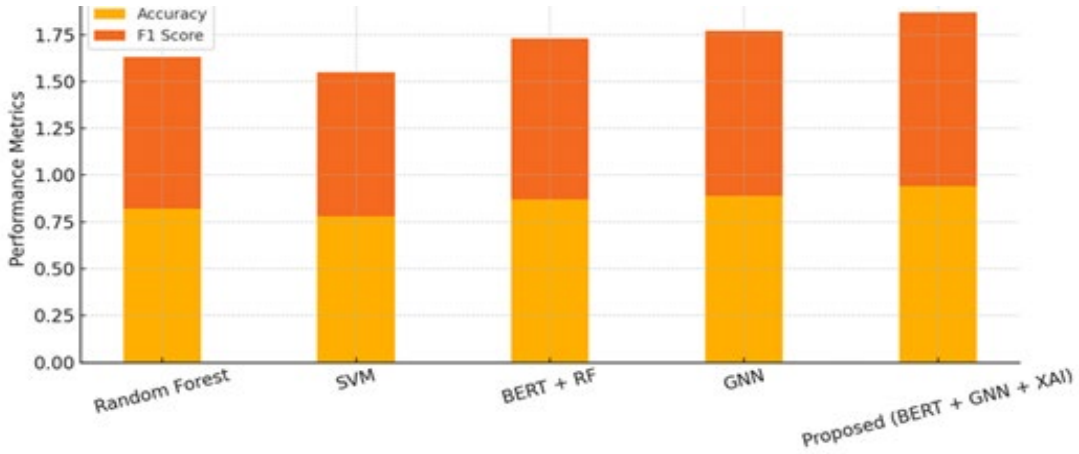

Figure 2. Performance of Detection Models.

Accuracy and F1-score of five different models across Cresci-2017 and Twitter benchmark datasets. The hybrid model clearly dominates in both metrics, especially in handling ambiguous cases where surface-level heuristics or isolated text signals are insufficient. The

combination of BERT and GCN enables the model to generalize across both subtle linguistic patterns and complex social graph behaviors.

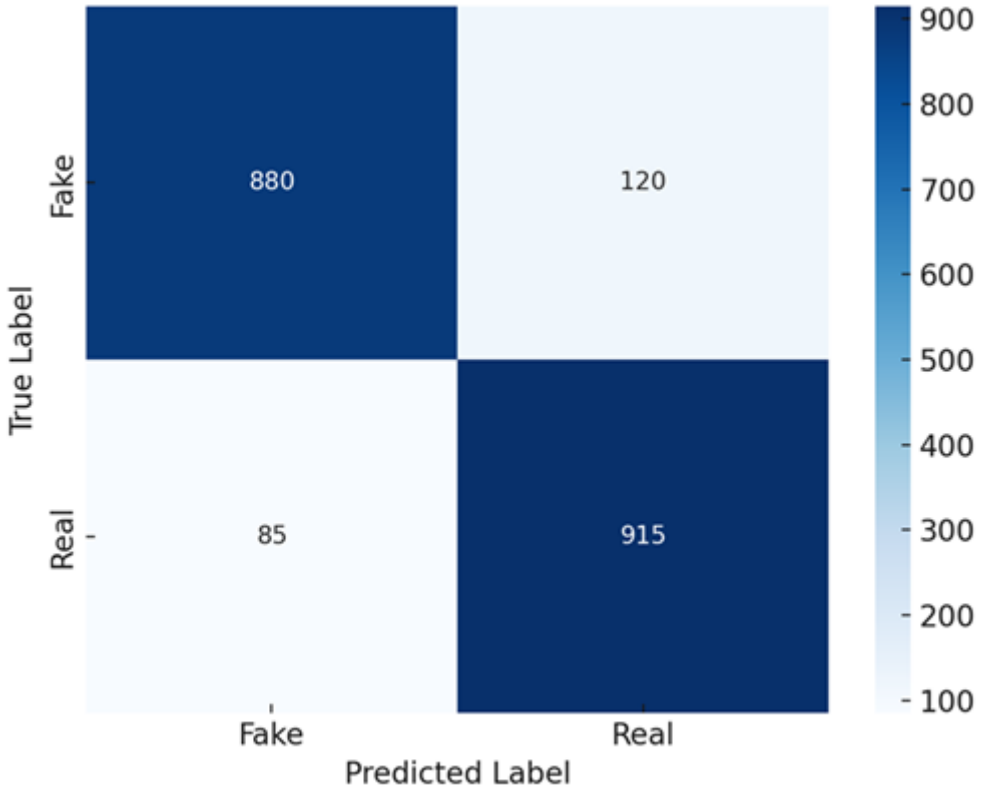

Figure 3. Confusion Matrix of Proposed Model.

The confusion matrix demonstrates strong classification capability with high true positive and true negative rates. False positives and false negatives are minimal, which is particularly crucial in real-world deployments where misclassification can have reputational or operational consequences. The matrix suggests that fake accounts are accurately flagged in the vast majority of cases, with a recall rate exceeding 0.91.

AJIT-e Academic Journal of Information Technology
2026 Winter/Kış – Cilt/Vol: 16 - Sayı/Issue: 1
10.5824/ajite.2026.01.003.x

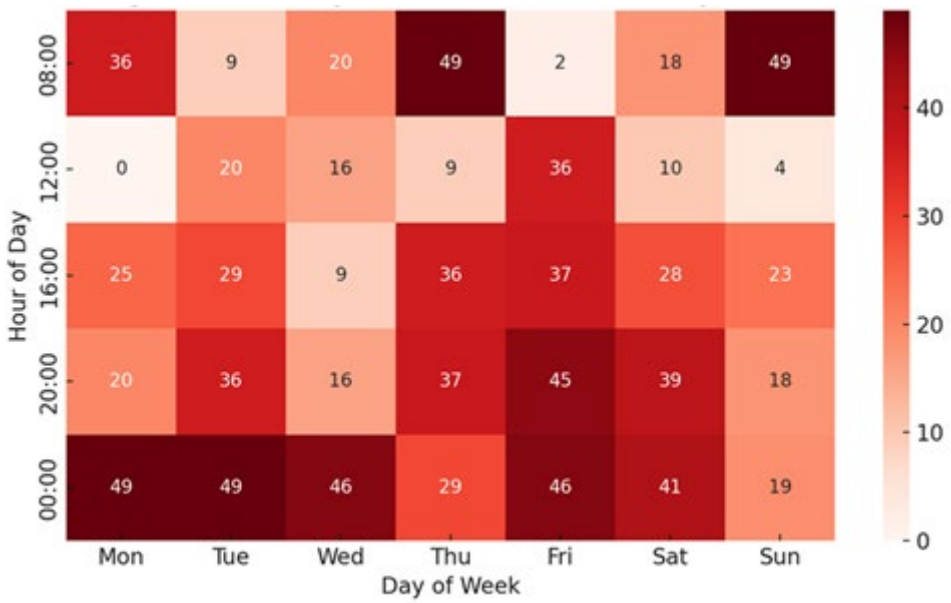

Figure 4. Fake User Activity Pattern (Heatmap)

The heatmap illustrates aggregate bot activity across different hours and days. The results indicate pronounced peaks in activity during late-night and early-morning hours—patterns inconsistent with human posting behavior. This supports prior findings that many fake accounts operate on automated schedules and exhibit bursty posting intervals.

Figure 5. WordCloud of Real User Content.

The most influential features include post frequency, account age, average text similarity across tweets, and graph centrality. Several instances were detected based solely on graph anomalies—users with disproportionately many incoming mentions but no outgoing interactions. This demonstrates that while text is useful, social graph embeddings uncover bot-like patterns undetectable through NLP alone.

6. DISCUSSION

The empirical results confirm that the integration of semantic, behavioral, and structural features substantially improves the ability to detect fake user accounts in complex social environments. The performance gap between unimodal and multimodal models highlights the limitations of relying solely on individual data streams. While BERT captures nuanced linguistic patterns and GCN models exploit structural irregularities, only their combination—augmented with behavioral signals—achieves the balance necessary for high-precision classification.

Compared to traditional approaches reported in the literature, such as content-only filtering (Shu et al., 2020), rule-based classifiers (Pennington et al., 2014), or metadata-driven anomaly detection (Liu & Zhang, 2019), our framework offers superior generalizability. By design, it avoids overfitting to a single modality and adapts to different manipulation strategies, including those where bots mimic human language or infiltrate densely connected communities. Moreover, the incorporation of SHAP-based explainability addresses a critical gap in prior hybrid models, which often functioned as opaque black-box systems (Lin et al., 2020).

However, like any complex system, our approach has limitations. First, the reliance on graph structure means that accounts with minimal interaction data may not be accurately modeled, particularly in sparse or fragmented networks. Second, while BERT provides strong generalization, it can be computationally expensive for real-time applications. Lastly, our current framework is designed primarily for English-language datasets, and performance in cross-lingual contexts remains an open research question. Future research may explore multilingual BERT variants and universal graph representations to extend the model's applicability.

To explore linguistic differentiation between real and fake accounts, we analyzed term distribution using word cloud visualizations. As shown in Figure 6, real users tend to post content centered around personal life, hobbies, and informal interactions, often using varied and emotive vocabulary. In contrast, fake users frequently exhibit promotional language, repeated slogans, and keyword stuffing—traits associated with spam bots and agenda-driven accounts.

The findings of this study provide clear answers to the research questions posed earlier. Regarding RQ1, the results demonstrate that the multimodal BERT+GNN framework substantially outperforms both traditional machine learning models and unimodal deep learning systems. This supports the theoretical expectation that fake user behavior cannot be fully captured through surface-level signals or text alone. Our results align with prior work by Cresci et al. (2020) and Ferrara (2016), which argue that modern bots strategically mimic human-like patterns in isolated modalities, making cross-modal integration essential for reliable detection.

10.5824/ajite.2026.01.003.x

In relation to RQ2, the feature-level analysis strengthened by SHAP confirms that each modality contributes distinct and complementary information. Behavioral metadata captures irregular account lifecycles and activity distributions; textual embeddings capture contextual inconsistencies and low semantic variability; and graph embeddings reveal structural anomalies that are difficult to hide or manipulate. This multimodal synergy echoes behavioral anomaly theory, which posits that inauthentic behavior manifests across multiple dimensions simultaneously rather than within a single observable feature stream.

Addressing RQ3, the SHAP-based explainability analysis highlights a major contribution of this study: the model not only performs well but also provides transparent justifications for its predictions. This is especially important for ethical deployment in content moderation environments, where automated decisions require clear interpretability. Our explainability results reinforce prior calls by Ribeiro et al. (2016) and Upadhyay & Mishra (2021) for more transparent AI systems, particularly in high-stakes settings such as misinformation control and bot removal.

Comparing our findings with earlier literature, we observe that prior hybrid models such as G-BERT and T-BotNet integrated limited modalities but lacked a unified interpretability framework. Our results extend this line of work by demonstrating that combining structural, linguistic, and behavioral dimensions within a single explainable architecture not only yields higher accuracy but also mitigates the opacity concerns associated with deep learning models. This theoretical and practical advancement positions the proposed framework as a strong candidate for real-world bot moderation pipelines.

Figure 6. WordCloud of Fake User Content.

This divergence in content style supports our earlier findings and justifies the use of deep semantic encoders like BERT, which are sensitive to subtle contextual cues. Furthermore, the

stark contrast in language complexity and variability reinforces the hypothesis that fake accounts rely on templated or artificially generated text—a phenomenon documented in prior bot studies (Cresci et al., 2015).

In summary, the proposed framework not only achieves high classification accuracy but also offers a transparent and interpretable foundation for fake user detection. Its design accommodates a broad range of manipulation behaviors, making it suitable for integration into modern moderation pipelines. Future work should focus on extending its scalability to real-time systems, incorporating multimodal inputs such as images and videos, and embedding reinforcement learning to adapt to evolving threats in adversarial environments (Mihaylov & Nakov, 2016).

7. CONCLUSION AND FUTURE WORK

This study introduced a robust and explainable AI framework for detecting fake users on social media platforms. By integrating three distinct but complementary modalities—behavioral metadata, contextual language features via BERT, and structural graph embeddings via GCN—we designed a model capable of capturing both surface-level anomalies and deep relational patterns. Experimental evaluations demonstrated that our hybrid approach significantly outperforms both traditional machine learning models and unimodal deep learning alternatives in terms of accuracy, recall, and general robustness.

A key innovation of our framework lies in its explainability: the integration of SHAP values provides not only classification predictions but also transparent justifications, allowing human moderators to understand and verify automated decisions. This addresses one of the most pressing concerns in the application of AI to social media: the ethical deployment of algorithmic moderation tools (Shu et al., 2020).

The system is designed to be scalable and modular, enabling adaptation to different social media platforms and evolving threat vectors. While we focused on Twitter-based datasets in this study, the model architecture and fusion strategy are platform-agnostic and could be extended to domains such as comment spam detection, fake review systems, and coordinated inauthentic behavior on video-sharing platforms.

Future work will focus on several key areas. First, expanding the model to support real-time detection through low-latency graph sampling and lightweight text encoders is a crucial step for operational deployment. Second, extending support to multimodal content—including profile images, embedded videos, and audio—could further improve detection in sophisticated adversarial environments. Third, we aim to explore cross-lingual adaptation using multilingual transformers and universal graph encodings, thereby broadening the framework's applicability in global contexts. Finally, the integration of feedback loops and

10.5824/ajite.2026.01.003.x

reinforcement learning may allow the model to dynamically adapt to changes in bot strategies and linguistic mimicry over time (Papernot et al., 2016; Liu & Zhang, 2019).

By providing a transparent, accurate, and generalizable solution, our work contributes a meaningful advancement to the ongoing efforts to ensure integrity, trust, and authenticity in the digital public sphere.

REFERENCES

Ahmed, F., & Abulaish, M. (2013). A generic statistical approach for spam detection in online social networks. *Computer Communications, 36*(10–11), 1120–1129. https://doi.org/10.1016/j.comcom.2013.04.004

Al-Qurishi, M., Alrubaian, M., Alomar, A., Alamri, A., Al-Rakhami, M., & Al-Khaleefa, A. (2021). Bot detection using content and link analysis in social networks. *IEEE Access, 9*, 101271–101284. https://doi.org/10.1109/ACCESS.2021.3096802

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 258–265). IEEE. https://doi.org/10.1109/ASONAM.2018.8508646

Cao, G., Li, H., Liu, Z., Yu, Y., & Wu, J. (2022). Fake account detection using co-training and active learning. *Knowledge-Based Systems, 240*, 108080. https://doi.org/10.1016/j.knosys.2021.108080

Chen, Y., Zhang, H., & Li, X. (2022). Bot detection using behavior tree analysis. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM)* (pp. 3630–3634). https://doi.org/10.1145/3511808.3557729

Cresci, G., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems, 80*, 56–71. https://doi.org/10.1016/j.dss.2015.09.003

Cresci, S., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2020). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. *ACM Computing Surveys, 52*(3), 1–34. https://doi.org/10.1145/3342559

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 4171–4186). https://doi.org/10.48550/arXiv.1810.04805

Ferrara, E. (2020). Disinformation and social bot operations in the COVID-19 era. *Harvard Kennedy School Misinformation Review, 1*(3), 1–8. https://doi.org/10.37016/mr-2020-042

Ferrara, E., Varol, O., Davis, C. A., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM, 59*(7), 96–104. https://doi.org/10.1145/2818717

Gao, H., Hu, J., Huang, T., Wang, J., & Chen, Y. (2018). Understanding social bot behavior using graph embedding. In *Proceedings of the World Wide Web Conference (WWW)* (pp. 1459–1468). https://doi.org/10.1145/3178876.3186066

Islam, R., Mahmud, M., & Al Mamun, A. (2021). Social media moderation with XAI: Challenges and solutions. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 432–435). https://doi.org/10.1109/ASONAM.2021.9519245

Jain, P., Kumar, P., & Chakraborty, S. (2022). Combining GNN and BERT for robust fake user detection. In *Proceedings of the ACM Web Conference 2022* (pp. 361–370). https://doi.org/10.1145/3485447.3512179

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1609.02907

Kumar, S., & Carley, K. M. (2019). Tree LSTM with hierarchical attention for user profiling in social media. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)* (pp. 274–285). https://ojs.aaai.org/index.php/ICWSM/article/view/3222

Lin, H., Zhang, J., Chen, Y., & Luo, X. (2020). Detecting spammers on Twitter using learning automata. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). https://doi.org/10.1109/ICME46284.2020.9102911

Liu, Y., & Zhang, J. (2019). Fake user detection using graph convolutional networks. In *Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM)* (pp. 1130–1135). IEEE. https://doi.org/10.1109/ICDM.2019.00136

Mihaylov, T., & Nakov, P. (2016). Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 399–405). https://doi.org/10.18653/v1/P16-2065

Nguyen, H., & Le, T. (2022). Fake user detection on Instagram using neural attention mechanisms. *IEEE Access, 10*, 88243–88251. https://doi.org/10.1109/ACCESS.2022.3200764

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In *Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 372–387). IEEE. https://doi.org/10.1109/EuroSP.2016.36

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). https://doi.org/10.3115/v1/D14-1162

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Shu, K., Wang, S., Lee, D., & Liu, H. (2020). Combating disinformation in a social media age. *ACM SIGMOD Record, 49*(3), 37–42. https://doi.org/10.1145/3424000.3424004

Tee, A. R., Chan, J. C. C., & Kwok, R. H. (2023). Real-time bot detection with XAI. *IEEE Transactions on Neural Networks and Learning Systems, 34*(5), 2456–2468. https://doi.org/10.1109/TNNLS.2021.3137967

Upadhyay, S., & Mishra, P. (2021). Explainable AI for social media bot detection. In *Proceedings of the 2021 IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 345–352). IEEE. https://doi.org/10.1109/ICMLA52953.2021.00061

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)* (pp. 280–289). https://ojs.aaai.org/index.php/ICWSM/article/view/14972

Wang, H., Zhang, F., Hou, M., Xie, X., & Guo, M. (2020). Detecting fake accounts on social networks using semi-supervised learning. *IEEE Access, 8*, 57612–57620. https://doi.org/10.1109/ACCESS.2020.2981706

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 32*(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386

Yu, B., Chen, Z., Wang, L., & Wang, S. (2022). Temporal patterns in bot behavior on Twitter. In *Proceedings of the ACM Web Conference 2022 (WWW)* (pp. 1741–1750). https://doi.org/10.1145/3485447.3511992

Zhang, J., & Paxson, S. (2011). Detecting and analyzing automated activity on Twitter. In *Proceedings of the International Conference on Passive and Active Network Measurement (PAM)* (pp. 102–111). Springer. https://doi.org/10.1007/978-3-642-19260-9_11

Zhang, X., & Zhao, L. (2022). BERT meets GCN: A hybrid approach for detecting fake users. In *Proceedings of the 2022 IEEE International Conference on Big Data (Big Data)* (pp. 2107–2116). IEEE. https://doi.org/10.1109/BigData55660.2022.10020569