SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Multimodal Framework for Explainable Fake-Profile Detection in Online Social Networks

Sudipta Acharjee[1], Dr. Pankaj Lathar[2]

[1]Research Scholar, Department of Computer Science and Applications Delhi Skill and Entrepreneurship University, New Delhi, India.
E-mail: sudipta.acharjee@dseu.ac.in
[2]Associate Professor, Department of Computer Science and Applications Delhi Skill and Entrepreneurship University, New Delhi, India.
E-mail: pankaj.lathar@dseu.ac.in

Abstract
Fake profiles and social bots increasingly combine plausible text, convincing images, regular-looking activity and carefully chosen social connections. A detector that examines only one source of evidence can therefore miss accounts that are deliberately designed to look ordinary. This paper presents a multimodal framework that combines four complementary evidence streams: semantic content, visual information, temporal behaviour, and social-graph structure. Transformer-based text representations, visual encoders, temporal activity features and graph neural representations are projected into a shared latent space. A profile-specific attention mechanism then assigns weights to the available modalities, while an explicit availability mask distinguishes missing evidence from genuinely low activity. The framework is designed for explainable decision support rather than an opaque binary verdict; modality contributions, feature attributions and neighbourhood evidence are intended to support analyst review. The supplied manuscript reports aggregate accuracy values of 85.2% for SVM, 88.6% for Random Forest, 91.4% for CNN, 93.8% for GNN and 96.7% for the proposed multimodal framework. These values are retained as reported evidence and are not presented here as independently reproduced experiments. The principal contribution of this manuscript is the integrated architecture, the missing-modality design, and a publication-oriented validation protocol that explicitly addresses ablation, calibration, temporal drift, cross-platform transfer, adversarial manipulation, and reproducibility.

Keywords: fake profiles; social bots; multimodal learning; graph neural networks; explainable AI; social media security; attention

1. Introduction

Online social networks have become an important part of everyday communication, professional networking, commerce and public discussion. Their openness also enables deceptive accounts to be created and operated at scale. Such accounts may be used for spam, impersonation, coordinated manipulation, fraud, misinformation, or other unwanted activity. Research on social bots has repeatedly shown that detection is an arms race: as platforms and researchers improve their detectors, account operators adapt their behaviour and appearance [2,7,16].

The challenge is no longer limited to obvious automation. A suspicious account may use fluent language, a plausible profile image, a conventional posting rhythm and a social neighbourhood that looks normal when each signal is inspected separately. Benchmark resources such as TwiBot-20 and TwiBot-22 were developed to support richer evaluation using user attributes, content and graph information [5,6]. More recent work has continued this movement toward multimodal and heterogeneous graph models [1,11].

This development motivates a multimodal view of profile authenticity. Text tells us something about what an account communicates; visual evidence tells us what it presents; behavioural evidence captures how and when it acts; and graph evidence describes how it interacts with other accounts. These signals are not interchangeable, and their reliability can vary from one profile to another. A useful fusion mechanism should therefore use available evidence without treating missing information as negative evidence.

The present work develops that idea into a four-branch framework with adaptive attention fusion. The intention is not to claim that multimodality, attention or graph learning is individually new; these directions are established in the literature [1,4,11,15,17]. Instead, the manuscript focuses on a coherent architecture, explicit handling of missing modalities, explainable decision support and a validation plan that separates reported benchmark values from results that still require independent reproduction.

1.1. Problem formulation

Let a profile p be represented by $X = \{X_S, X_v, X_b, X_g\}$, where X_S is semantic content, X_v is visual evidence, X_b is behavioural evidence, and X_g is graph or relational information. The task is to estimate $y \in \{0,1\}$, where y denotes

the authenticity class. The detector is expressed as y-hat = f(X_s, X_v, X_b, X_g). In practice, one or more components may be unavailable, noisy or deliberately manipulated. The design goal is therefore not merely to maximise classification accuracy, but to combine corroborating evidence while accounting for each modality's reliability and availability.

1.2. Research questions

- RQ1. Does multimodal fusion provide a measurable advantage over single-modality and conventional baselines?
- RQ2. Which evidence streams contribute most strongly to the final decision?
- RQ3. Does adaptive fusion remain useful when one or more modalities are missing or noisy?
- RQ4. How stable is the detector under temporal drift and cross-platform transfer?
- RQ5. Can modality-level explanations and calibrated confidence improve analyst review without being interpreted as causal proof?

1.3. Contributions

- A unified four-branch architecture covering semantic, visual, behavioural and graph evidence.
- An adaptive fusion mechanism that learns profile-specific modality weights rather than fixing them manually.
- An explicit availability mask so missing evidence is distinguished from weak evidence.
- An explainability layer that combines modality evidence, feature attribution and graph context while avoiding causal over-interpretation.
- A validation protocol covering ablation, repeated trials, calibration, temporal generalisation, cross-platform evaluation and adversarial robustness.

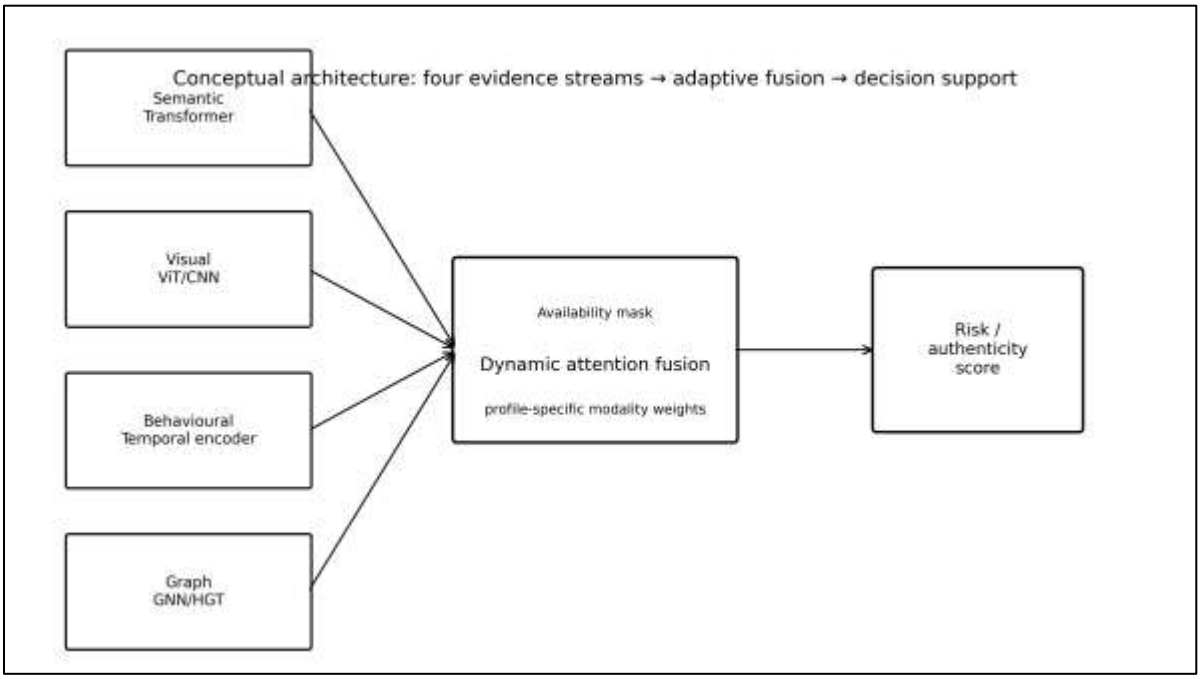

Fig. 1. Conceptual architecture of the proposed multimodal detector. The schematic describes the method and does not represent newly measured experimental data.

2. Related Work / Literature Review

Early social-bot and fake-account research relied on metadata, activity patterns and manually designed behavioural indicators. These approaches remain attractive because they are comparatively inexpensive to compute, but sophisticated accounts can imitate ordinary behaviour and reduce the value of any single handcrafted signal [2,7,16]. Deep representation learning has since enabled richer content- and user-level representations. Kudugunta and Ferrara, for example, combined textual and user-level information in a neural bot-detection architecture [10].

Graph learning offers another important perspective because suspicious accounts often operate in groups rather than as isolated users. GraphSAGE established an inductive approach for generating node representations from local neighbourhoods [9], while graph attention networks introduced learnable weighting of neighbouring nodes [17]. Relational graph convolution has also been applied specifically to Twitter bot detection through BotRGCN, which combines heterogeneous graph structure with multimodal user information [18].

Multimodal and heterogeneous fusion has become increasingly relevant as adversaries combine content, behaviour and network camouflage. Recent work has explicitly fused heterogeneous features and social relationships, including a multimodal heterogeneous graph-transformer approach [11] and a multi-source heterogeneous feature framework [1]. The literature therefore points toward richer evidence integration, but it also suggests that benchmark composition, label quality, graph completeness, temporal coverage and class balance can materially affect reported performance [5,6].

The research gap addressed here is consequently methodological as well as architectural. A framework may be multimodal without being well validated. The present work combines four evidence families and explicitly models missing modalities. It defines an evaluation protocol that includes ablation, repeated trials, calibration, temporal and cross-platform evaluation, adversarial testing and reproducibility.

3. Methodology

The framework contains four parallel representation branches. Each branch converts a different evidence family into a latent vector. The vectors are normalised and projected to a common dimensionality before fusion. The final decision layer produces an authenticity probability or risk score. The architecture is modular: you can replace a branch with a stronger encoder without changing the overall fusion logic.

Table 1. Functional specification of the proposed multimodal architecture.

Module	Main inputs	Representation and role
Semantic	Biographies, posts, comments, captions, hashtags, usernames	Transformer embeddings and linguistic similarity, repetition and inconsistency features.
Visual	Profile images and shared media	CNN or vision-transformer embeddings plus image–text consistency signals.
Behavioural	Posting times, intervals, engagement and growth	Temporal features describing burstiness, repetition, regularity and coordination.
Graph	Follow, reply, mention, retweet, quote, hashtag and URL relations	GNN or heterogeneous-graph representation of neighbourhood and coordination structure.
Adaptive fusion	Four modality vectors + availability mask	Profile-specific attention weights and a shared fused representation.
Decision	Fused representation	Authenticity probability/risk score for analyst support.

3.1. Semantic representation

The semantic branch processes the text that is available for a profile. Transformer encoders are suitable because they model contextual relationships rather than treating words as independent tokens. BERT is a representative example [3], while the original Transformer introduced the self-attention mechanism used by many modern language models [15]. The semantic branch can also derive account-level signals such as duplicated biographies, repeated templates, unusual lexical regularity, semantic similarity across accounts, and mismatches between profile descriptions and recent posts.

For a transformer, attention can be written as Attention(Q,K,V) = softmax(QK^T / sqrt(d_k))V [15]. The equation specifies the representation mechanism rather than claiming a new attention operator.

3.2. Visual representation

The visual branch analyses profile photographs and other shared media. Vision Transformer architectures provide a patch-based image representation [4]. A further cross-modal signal can be obtained by comparing image and text embeddings; CLIP demonstrated the feasibility of learning a shared image–text representation space [13]. In fake-profile detection, however, visual similarity should be treated as supporting evidence. A reused or synthetic image is not, by itself, proof that the associated account is malicious.

3.3. Behavioural representation

Behavioural evidence captures the rhythm and regularity of account activity. Candidate variables include posting frequency, inter-event intervals, active time windows, reply-to-retweet ratios, follower and following growth, engagement regularity and repeated content. The purpose is to identify patterns that are difficult to understand from a static profile alone. No individual behavioural feature is assumed to be conclusive.

3.4. Graph representation

The graph is defined as G=(V, E), where nodes represent accounts or related entities and edges represent relations such as follows: replies, mentions, retweets, quotes, shared hashtags or shared URLs. Graph-based representations are valuable because coordinated accounts can appear ordinary when examined individually. GraphSAGE provides inductive neighbourhood aggregation [9], graph attention networks provide learnable neighbourhood weighting [17], and BotRGCN uses relational graph convolution with multimodal user information for Twitter bot detection [18].

3.5. Dynamic attention fusion

Let h_s, h_v, h_b and h_g denote the projected semantic, visual, behavioural and graph representations. The fused representation is $H = \Sigma_i \alpha_i h_i$, subject to $\Sigma_i \alpha_i = 1$. The weights α_i are learned from the available evidence. A binary availability mask m_i is supplied to the fusion network so that an unavailable modality is not interpreted as evidence of low activity. This distinction matters in practical social-media settings, where API restrictions, deleted content and privacy controls can remove otherwise useful observations.

The design can also use modality dropout during training. The model is periodically exposed to incomplete inputs so that the decision layer learns not to depend on one evidence stream. The resulting architecture is intended to degrade gracefully rather than fail when a single modality is absent.

3.6. Explainability and operational interpretation

The framework treats explanation as part of the interface between the model and the analyst. Modality weights can indicate which evidence family influenced a prediction, while feature-level attribution can identify local factors that contributed to a particular score. SHAP provides a general framework for feature attribution in complex models [12]. An attention weight should not be described as proof that a feature caused a profile to be fake; explanations should instead be evaluated for fidelity, stability and usefulness.

3.7. Robustness, privacy and ethics

Adversarial adaptation is a central concern. An attacker can paraphrase text, replace a profile image, change posting times, alter interaction patterns or modify the social neighbourhood. Multimodal fusion provides redundancy, but redundancy should not be confused with immunity. The proposed evaluation therefore includes controlled perturbations and realistic threat models.

Privacy is equally important because the evidence streams include potentially sensitive information about communications, images, activity and relationships. Data minimisation, access controls, retention limits and audit trails should be part of deployment. Where graph data are distributed across organisations, federated approaches may reduce centralisation but introduce additional concerns such as non-IID data, poisoned updates and information leakage. Privacy should therefore be treated as a system-level security requirement.

4. Dataset and Preprocessing

4.1. Reported corpus

The supplied manuscript describes a corpus of 100,000 profiles comprising 50,000 genuine profiles, 40,000 fake profiles and 10,000 adversarial synthetic profiles. These counts are retained because they are part of the supplied manuscript. This revision does not claim that we independently reconstructed the corpus. Before submission, the exact source datasets, versions, inclusion criteria, label definitions, synthetic-profile generation procedure and deduplication policy should be documented so that the experiment can be audited.

Table 2. Corpus composition reported in the supplied manuscript; provenance remains to be documented for independent reproduction.

Profile category	Reported count	Share
Genuine profiles	50,000	50%
Fake profiles	40,000	40%
Adversarial synthetic profiles	10,000	10%
Total	100,000	100%

4.2. Preprocessing

Text processing should include normalisation, tokenisation and platform-specific noise removal while preserving information relevant to authenticity. Resize and normalise images using a fixed pipeline. Order behavioural events by time and transform them into comparable temporal features. Construct the graph after account-level deduplication and identity-leakage checks.

4.3. Data split and leakage control

Separate training, validation, and test sets at the account level. Near-duplicate accounts, copied content, and repeated images should not cross the split. Where temporal information is available, report an additional chronological split to evaluate performance under distribution shift. These controls are essential because random row-level splitting can produce overly optimistic estimates when related profiles or copied content appear in both training and test sets.

5. Experimental Setup / Baselines

5.1. Evaluation measures

Accuracy alone is not sufficient for an authenticity detector, particularly when class proportions vary. The evaluation protocol therefore calls for precision, recall, F1, macro-F1, ROC-AUC and PR-AUC, together with specificity, sensitivity, Matthews correlation coefficient and calibration error where appropriate. Accuracy, precision, recall and F1 are defined as follows: Accuracy=(TP+TN)/(TP+TN+FP+FN); Precision=TP/(TP+FP); Recall=TP/(TP+FN); and F1=2PR/(P+R).

Calibration deserves explicit attention because an analyst may use predicted probability to prioritise review. Modern neural networks can be poorly calibrated even when classification accuracy is strong; temperature scaling is one established post-hoc approach [8]. The final empirical study should therefore report a calibration metric alongside discrimination metrics.

5.2. Baseline models

The supplied manuscript compares SVM, Random Forest, CNN and GNN baselines. These provide a useful progression from classical learning to deep and graph-based models. A journal-grade comparison should additionally include strong contemporary multimodal and heterogeneous-graph baselines where the dataset permits. Recent work on heterogeneous feature fusion, graph transformers and relational graph models provides appropriate reference points [1,11,18].

5.3. Ablation and repeated trials

A complete ablation should include each modality individually, pairwise combinations, three-modality combinations, four-way fusion without attention and four-way fusion with adaptive attention. Missing-modality tests should remove text, images, behaviour and graph information separately. Repeated runs with multiple random seeds should be used to report mean and standard deviation rather than a single favourable run.

5.4. Reproducibility protocol

The final experimental record should identify dataset versions, label definitions, split rules, preprocessing operations, hyperparameters, random seeds, software libraries, hardware, training time and model checkpoints. The code should reproduce the tables and figures from a clean environment. If platform data cannot be legally redistributed, provide persistent dataset identifiers and a lawful reconstruction procedure.

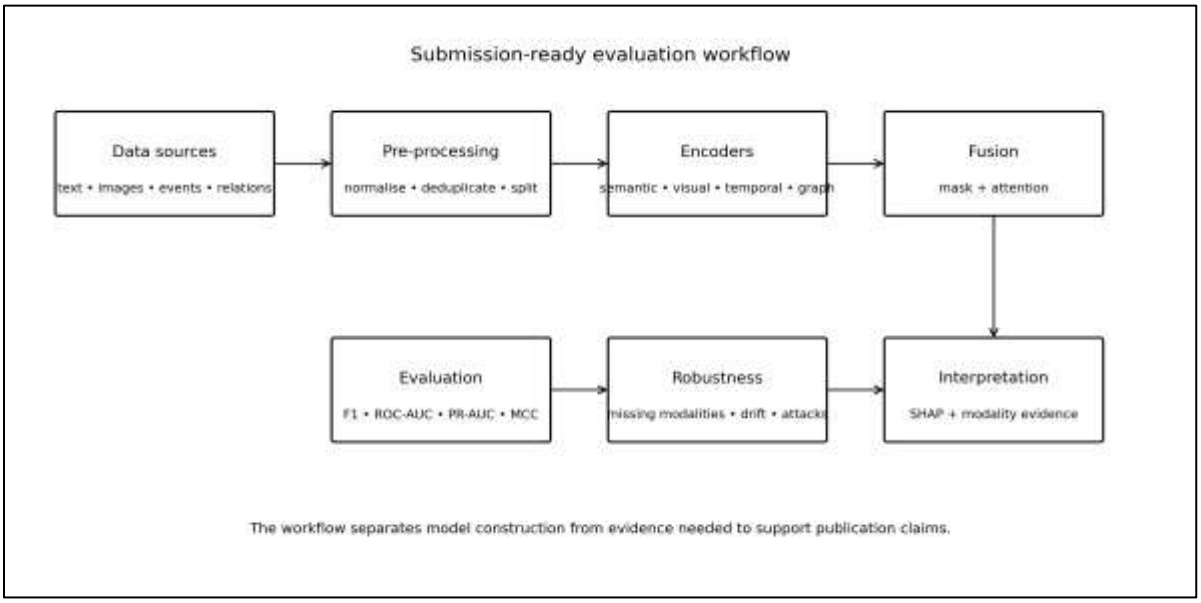

Fig. 2. Evaluation-oriented workflow from multimodal inputs through representation learning, robustness testing and explainable decision support.

6. Results and Discussion

6.1. Reported benchmark comparison

We reproduce the numerical comparison in the manuscript here, without implying that we performed new experiments during this revision. The reported values are 85.2% for SVM, 88.6% for Random Forest, 91.4% for CNN, 93.8% for GNN and 96.7% for the proposed multimodal framework. These values therefore serve as a benchmark illustration and reflect what is reported in the source manuscript, not independently verified measurements.

Table 3. Aggregate accuracy values retained from the supplied manuscript; no independent reproduction is claimed.

Model	Reported accuracy	Difference from SVM
SVM	85.2%	—
Random Forest	88.6%	+3.4 percentage points
CNN	91.4%	+6.2 percentage points
GNN	93.8%	+8.6 percentage points
Proposed multimodal framework	96.7%	+11.5 percentage points

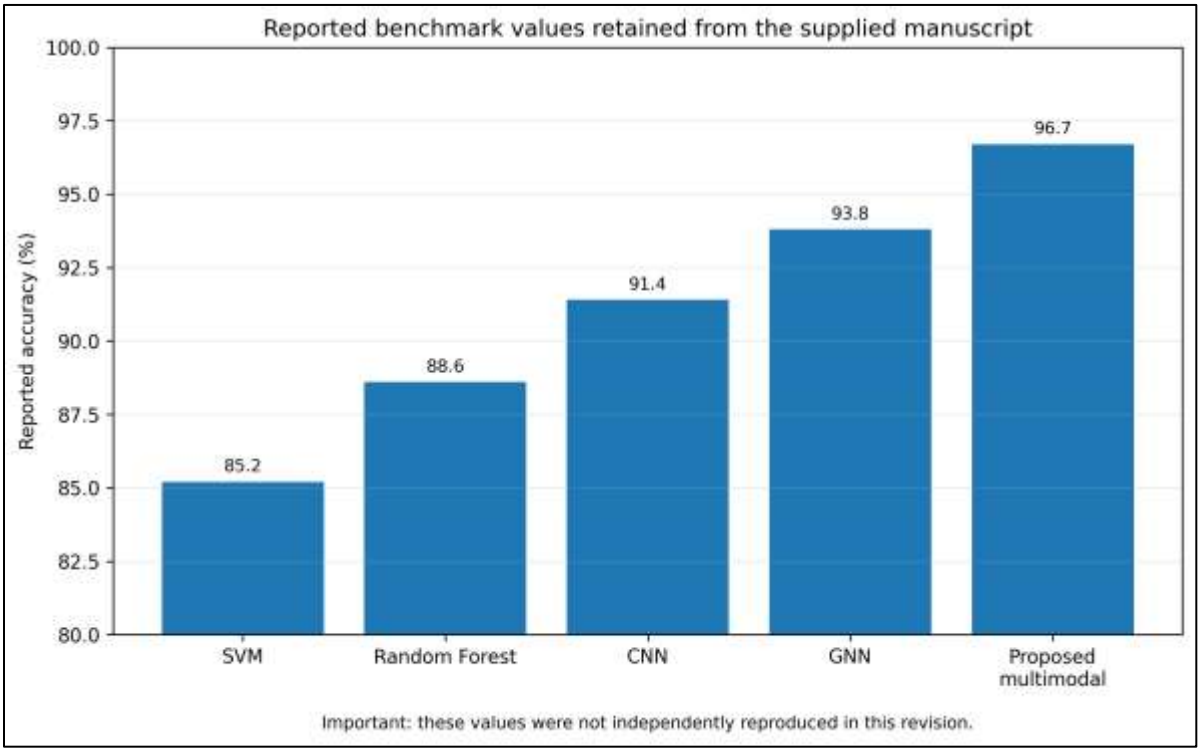

Fig. 3. Reported accuracy values retained from the supplied manuscript. The chart visualises supplied values and is not a newly measured performance curve.

6.2. Interpretation

The aggregate comparison is directionally consistent with the central design hypothesis: combining different evidence families can provide more information than relying on one family alone. At the same time, the difference between 96.7% and a baseline cannot be interpreted as statistically significant without repeated runs, uncertainty estimates and appropriate paired tests. Nor can accuracy establish robustness to distribution shift, missing modalities or adversarial manipulation.

6.3. Evidence required for a definitive performance claim

- Report precision, recall, F1, macro-F1, ROC-AUC and PR-AUC for every baseline and the proposed model.
- Provide class-wise confusion matrices and uncertainty intervals.
- Repeat training across multiple random seeds and report mean ± standard deviation.
- Perform complete modality ablation and missing-modality evaluation.
- Test temporal generalisation by training on earlier data and evaluating on later data.
- Evaluate cross-platform transfer when compatible datasets are available.

• Define and reproduce adversarial threat models covering text, images and graph relations.
• Report inference latency, memory use, training cost and scalability characteristics.

6.4. Practical interpretation

A fake-profile detector is best understood as decision support rather than a final authority. A practical analyst-facing output could contain the risk score, the most influential evidence categories, selected supporting features, relevant graph neighbours and a confidence or calibration indicator. Human review is particularly important when false positives could have significant consequences.

7. Limitations / Threats to Validity

• The 100,000-profile corpus described in the supplied manuscript has not been independently reconstructed in this revision.
• The provenance, label-generation procedure and synthetic-adversarial construction require fuller documentation.
• Complete class-wise metrics, confidence intervals, or calibration results do not accompany the reported accuracy values.
• Full hyperparameters, exact data splits, random seeds and hardware/software records are not available in the supplied manuscript.
• Temporal and cross-platform generalisation have not yet been demonstrated.
• Adversarial robustness remains a proposed evaluation dimension rather than a completed experiment.
• Platform API restrictions and privacy policies may limit reproducibility of some multimodal evidence.
• The reported aggregate values should not be interpreted as evidence of universal superiority until independent reproduction is completed.

These limitations are not treated as minor editorial issues. They define the boundary between a plausible methodological framework and a fully validated empirical contribution. The final scientific claim should therefore remain proportional to the evidence actually reproduced.

8. Conclusion

This paper presents a multimodal framework for fake-profile detection in online social networks by combining semantic, visual, behavioural and graph evidence through adaptive attention fusion. An explicit missing-modality mask allows the system to distinguish unavailable evidence from genuinely weak activity, while the explainability layer is intended to support analyst review rather than replace it. This manuscript reports 96.7% accuracy for the proposed approach, compared with 85.2%, 88.6%, 91.4% and 93.8% for the listed baselines. These figures are retained as reported benchmark evidence, not as independently reproduced measurements. The paper's strongest contribution is therefore the integrated methodology and its explicit validation framework. Before making a definitive superiority claim, the study should establish reproducibility through complete ablation, repeated trials, calibration, temporal and cross-platform testing, and realistic adversarial evaluation.

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data availability

We have used the Twitter Fake Account Dataset, TwiBot-20, the Cresci fake-profile resources, and FakeNewsNet as relevant dataset resources downloaded from the Kaggle Website. It refers to a CSV file named fake_profiles.csv containing fields such as followers, following, posts, biography, and label. The final submission should state the exact dataset versions and subsets used, along with label definitions, inclusion/exclusion criteria, preprocessing, and split rules. If platform-derived data cannot legally be redistributed, provide persistent identifiers and a lawful reconstruction procedure.

Code availability

The supplied manuscript describes implementations for metadata preprocessing, TF-IDF text features, feature combination, Random Forest classification, an LSTM text model and a Flask prediction API, together with possible cloud mappings. For the final empirical paper, the complete preprocessing, model training, evaluation, ablation and statistical-analysis code should be deposited in a version-controlled repository with dependency specifications, random seeds and instructions for reproducing the tables and figures.

Declaration of generative AI and AI-assisted technologies

During the preparation of this revised manuscript, the authors used ChatGPT (OpenAI) with ethics as an AI-assisted tool for manuscript organisation, language refinement, clarity and restructuring. The authors are responsible for critically reviewing, verifying and editing the material, including all scientific claims, interpretations and references, and retain full responsibility for the final published content.

Declaration of competing interests
The authors declare no competing financial or personal interests that could have influenced the work.

Credit Authorship contribution statement
Sudipta Acharjee: Conceptualisation, Methodology, Investigation, Writing – original draft, Writing – review & editing, Visualisation. Pankaj Lathar: Supervision, Validation, Writing – review & editing.

Ethics statement
The study analyses online-account information for research on computational authenticity detection. This submission includes a statement of the applicable dataset licences, privacy safeguards, institutional approval status, and any restrictions governing the use or redistribution of platform-derived data.

References

[1] Cheng, M., Xiao, Y., Huang, T., Lei, C., & Zhang, C. (2025). CB-MTE: Social bot detection via multi-source heterogeneous feature fusion. Sensors, 25(11), 3549. https://doi.org/10.3390/s25113549.
[2] Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). The paradigm shift of social spambots: Evidence, theories, and tools for the arms race. In WWW '17 Companion, 963–972. https://doi.org/10.1145/3041021.3055135.
[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, 4171–4186. https://doi.org/10.18653/v1/N19-1423.
[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.
[5] Feng, S., Wan, H., Wang, N., Li, J., & Luo, M. (2021). TwiBot-20: A comprehensive Twitter bot detection benchmark. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management. https://doi.org/10.1145/3459637.3482019.
[6] Feng, S., Tan, Z., Wan, H., Wang, N., Chen, Z., Zhang, B., Zheng, Q., Zhang, W., Lei, Z., Yang, S., et al. (2022). TwiBot-22: Towards graph-based Twitter bot detection. Advances in Neural Information Processing Systems, 35, 30856–30869.
[7] Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. Communications of the ACM, 59(7), 96–104. https://doi.org/10.1145/2818717.
[8] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, 70, 1321–1330.
[9] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems, 30.
[10] Kudugunta, S., & Ferrara, E. (2018). Deep neural networks for bot detection. Information Sciences, 467, 312–322. https://doi.org/10.1016/j.ins.2018.08.019.
[11] Luo, J., & Jin, C. (2025). Fusing content and social relationships: A multi-modal heterogeneous graph transformer approach for social bot detection. EPJ Data Science, 14, 68. https://doi.org/10.1140/epjds/s13688-025-00583-5.
[12] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions in Advances in Neural Information Processing Systems, 30.
[13] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., et al. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, 139, 8748–8763.
[14] Sarfraz, A., Ahmad, A., Zeshan, F., Hamid, M., & Alshalali, T. A. N. (2025). Unmasking deception: Detection of fake profiles in online social ecosystems. Journal of Big Data, 12, 214. https://doi.org/10.1186/s40537-025-01254-y.
[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need.
[16] Varol, O., Ferrara, E., Davis, C., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterisation. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 280–289. https://doi.org/10.1609/icwsm.v11i1.14871.
[17] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In International Conference on Learning Representations.