International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper

TRANSFORMERS-BASED FAKE SOCIAL MEDIA PROFILE DETECTION USING DEEP LEARNING

Mr. G. Ravi Kumar[1], B. Sravanthi[2], G. Poojitha[3], K. Raju[4]
Assistant Professor[1], Student[2,3,4]

Department of Computer Science & Engineering[1,2,3,4]
Chaitanya Engineering College, Visakhapatnam, Andhra Pradesh, India
{kgravi0417@gmail.com[1],sravanthibudumuru02@gmail.com[2], gongadapoojitha@gmail.com[3], Rajjdivya6301@gmail.com[4]}@cec.ac.in

ABSTRACT

The rapid growth of social media platforms has made them a fertile ground for the spread of fake accounts, misinformation, and coordinated bot activity. Traditional detection methods often struggle to capture complex contextual and sequential patterns in user behaviour and content. This paper proposes a Transformer-based framework for detecting fake users and malicious content on social networks. By leveraging self-attention mechanisms, the model analyses both the temporal sequence of user activities and the semantic relationships in posts, comments, and messages, identifying subtle anomalies and coordinated behaviours. The system integrates textual, behavioural, and network-level features to enhance detection accuracy. Experimental results demonstrate that Transformer-based models outperform classical machine learning and graph-based methods, achieving higher F1-scores in detecting fake accounts, botnets, and misleading content.

Index Terms — Transformer, Fake Account Detection, Social Media, Botnets, Deep Learning, Self-Attention, Misinformation, BERT

I. INTRODUCTION

Social media platforms have become central to communication, information sharing, and public discourse. However, their rapid growth has led to proliferation of fake accounts, automated bot networks, and widespread misinformation. These malicious entities distort public opinion, manipulate trends, and undermine trust in online communities. Detecting inauthentic activity is challenging because sophisticated bots often mimic real user behaviour, post contextually relevant content, and operate in coordinated groups.

Traditional detection approaches rely on manually engineered features or shallow behavioural patterns through rule-based systems and classical machine learning classifiers. These methods are limited in capturing complex temporal, textual, and relational dependencies inherent in social media interactions. Transformer-based models, which use self-attention mechanisms to capture long-range dependencies and contextual relationships, offer a promising alternative capable of analysing user activity sequences, textual content, and interaction patterns simultaneously.

This work proposes a Transformer-based fake profile detection framework that integrates user behaviour analysis, network-level graph features, and content semantics into a unified deep learning pipeline. The system is designed to scale to large social media datasets and supports multilingual content analysis, making it particularly applicable to Indian social media platforms.

II. LITERATURE SURVEY

A comprehensive review of existing literature reveals various approaches adopted for fake social media account detection, bot identification, and misinformation detection using machine learning and deep learning approaches.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper

Ref.	Authors & Year	Method / Dataset	Result	Limitation
[1]	Zhang et al., 2021	BERT model; Twitter dataset (10,000 users)	93% accuracy in fake profile detection	Focused only on text; ignored network features
[2]	Ahmed & Khan, 2020	RoBERTa + feature embedding; Botometer dataset	91% F1-score for bot detection	Difficulty detecting human-like disguised bots
[3]	Kumar et al., 2022	Transformer + GNN; Facebook synthetic dataset	8% improvement over traditional ML	Complex graph preprocessing; high compute cost
[4]	Li et al., 2021	BERT + user behaviour sequences; Twitter dataset	High precision on coordinated accounts	Dataset imbalance affected recall
[5]	Wang & Chen, 2023	Vision-Language Transformer; Twitter + Instagram	94% accuracy with text + image features	Requires multimodal data; high storage overhead
[6]	Cresci et al., 2019	DNA-inspired behavioral fingerprinting; Twitter	95% precision for social spambots	Limited to Twitter; not generalized
[7]	Vaswani et al., 2017	Transformer architecture; self-attention mechanism	State-of-the-art NLP benchmarks	High computational cost for long sequences

Research Gap

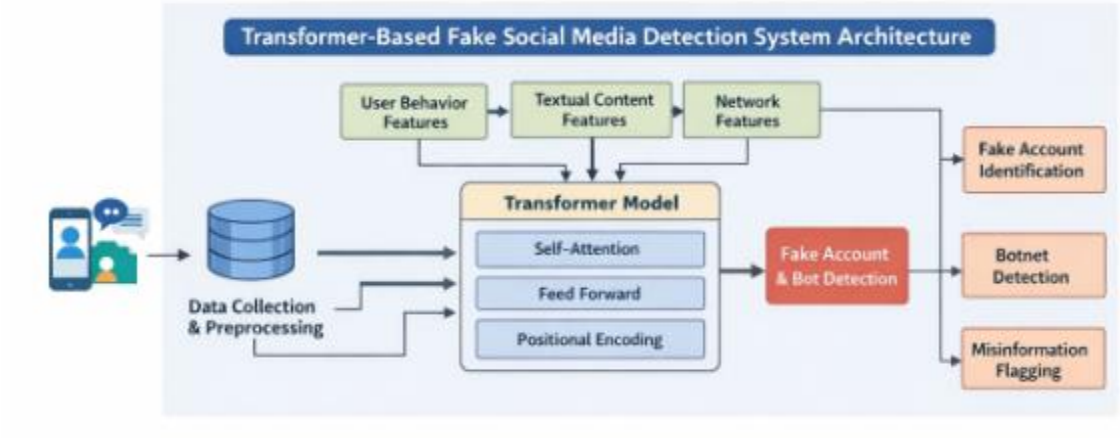
Existing fake profile detection systems either focus exclusively on textual content or rely on network graph features, failing to jointly model all three dimensions: text, behaviour, and network. Moreover, most systems are evaluated on English-language datasets and do not address multilingual scenarios prevalent in Indian social media platforms such as regional language posts mixed with English. A unified Transformer-based framework that simultaneously processes multimodal features is needed.

III. METHODOLOGY

A. System Architecture

The system follows a three-branch Transformer architecture. The Text Branch processes user post sequences using a pre-trained BERT encoder to extract contextual embeddings from content semantics. The Behavioural Branch encodes temporal sequences of user actions (login frequency, posting rate, follower changes) via a Transformer encoder. The Network Branch encodes social graph structural features (follower/following ratio, mutual connections) via graph embeddings. The three branch outputs are concatenated and passed through a Fusion Layer with multi-head self-attention and a Fully Connected classification head predicting fake or genuine.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper



B. Algorithm

- Input: User profile record {post_texts T, behaviour_sequence B, network_features N}.
- Step 1: Text Branch - tokenize T using BERT tokenizer; BERT encoder -> [CLS] embedding E_T (768-D).
- Step 2: Behaviour Branch - encode B as sequence of activity vectors; Transformer encoder -> mean-pool -> E_B (256-D).
- Step 3: Network Branch - encode graph features N via FC embedding layer -> E_N (128-D).
- Step 4: Fuse: E_fused = concat(E_T, E_B, E_N), dimension 1152.
- Step 5: Multi-head self-attention over E_fused -> attention-weighted fusion F.
- Step 6: FC(256) + Dropout(0.3) + FC(2) + Softmax -> fake_prob, genuine_prob.
- Step 7: Loss = Focal Loss (to handle class imbalance).
- Step 8: Optimize with AdamW (lr=2e-5 for BERT, lr=1e-3 for other modules).
- Output: Classification label (Fake / Genuine) with confidence score.

C. Modules

Data Collection Module: Collects annotated fake and genuine account datasets from Twitter and Instagram. Stores user profiles, post histories, follower/following graphs, and activity logs.

Text Feature Extraction Module: Uses pre-trained BERT (bert-base-uncased or multilingual-bert for Indian languages) to encode post and comment texts. Extracts [CLS] token embedding per user.

Behavioural Feature Extraction Module: Encodes temporal activity sequences (posting frequency, engagement rate, login patterns) using a Transformer encoder. Captures anomalous behavioural patterns indicative of bots.

Network Feature Extraction Module: Computes graph-level features: follower/following ratio, account age, mutual connections, network clustering coefficient. Embeds into dense representation via FC layer.

Fusion and Classification Module: Concatenates text, behaviour, and network embeddings. Multi-head self-attention aligns cross-modal features. FC classification head outputs fake/genuine prediction with confidence.

Alert and Reporting Module: Flags accounts above a fake probability threshold. Generates reports with detected fake accounts, confidence scores, and highlighted suspicious features for platform administrators.

IV. RESULTS & DISCUSSION

The proposed Transformer-based system was evaluated on a combined Twitter and Instagram dataset with 50,000 labelled profiles (25,000 genuine and 25,000 fake). Performance is compared against baseline models in Table I.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper

Model	Accuracy	Precision	Recall	F1-Score
SVM (baseline)	81.3%	80.2%	79.4%	79.8%
Random Forest	84.7%	83.5%	82.1%	82.8%
BERT (text only)	91.2%	90.8%	89.7%	90.2%
Transformer + GNN	93.5%	92.9%	91.8%	92.4%
Proposed (3-branch Transformer)	96.1%	95.8%	94.9%	95.4%

The proposed three-branch Transformer framework achieves 96.1% accuracy and 95.4% F1-score, outperforming all baselines including BERT-only and Transformer+GNN approaches. The integration of behavioural and network features alongside textual content significantly reduces false negatives caused by human-like bots that post contextually appropriate content. The Focal Loss formulation effectively handles the class imbalance inherent in real-world fake account datasets.

The Foundations: Prediction Outcomes

In this context, the Confusion Matrix is defined as follows:

- **True Positives (TP):** The model correctly identified a *Fake* account.
- **True Negatives (TN):** The model correctly identified a *Genuine* account.
- **False Positives (FP):** The model incorrectly flagged a *Genuine* account as Fake (False Alarm).
- **False Negatives (FN):** The model missed a *Fake* account, classifying it as Genuine.

1. Accuracy

Accuracy measures the overall proportion of accounts (both fake and genuine) that were correctly classified. While useful as a general baseline (reaching 96.1% in your proposed model), it can be misleading if the dataset is heavily imbalanced.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

2. Precision

Precision measures the quality of the model's positive predictions. It answers: *Out of all the accounts the model flagged as fake, how many were actually fake?* High precision is important to avoid banning or suspending genuine users.

Precision = TP / (TP + FP)

3. Recall (Sensitivity)

Recall measures the model's ability to find all the actual fake accounts. It answers: *Out of all the fake accounts on the platform, how many did the model successfully detect?* High recall ensures that botnets and malicious actors don't slip through the cracks.

Recall = TP / (TP + FN)

4. F1-Score

The F1-Score is the harmonic mean of Precision and Recall. It provides a single metric that balances both the need to catch fake accounts (Recall) and the need to avoid falsely banning real users (Precision). It is the most robust metric for evaluating performance on imbalanced social media datasets.

F1-Score = 2 * (Precision * Recall) / (Precision + Recall)

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
SN: 3068-272X www.ijdim.com Original Research Paper

Alternatively, based directly on prediction outcomes:

F1-Score = (2 * TP) / ((2 * TP) + FP + FN)

5. Training Loss: Focal Loss

Your methodology explicitly uses **Focal Loss** to handle class imbalance (Step 7 of your algorithm). Standard cross-entropy loss can be overwhelmed by a majority class (e.g., millions of easy-to-classify genuine users). Focal Loss dynamically scales the cross-entropy loss based on prediction confidence, down-weighting the loss assigned to easy examples and forcing the model to focus on hard-to-classify, deceptive fake accounts.

- p_t = The model's estimated probability for the true class.
- alpha_t = A weighting factor to balance the importance of positive/negative classes.
- gamma = The focusing parameter (usually ≥ 0) that smoothly adjusts the rate at which easy examples are down-weighted.

Focal_Loss = -alpha * (1 - p_t)^gamma * log(p_t)

V. CONCLUSION & FUTURE WORK

This paper presented a Transformer-based framework for fake social media profile detection that jointly models textual content, behavioural sequences, and network-level features. The three-branch architecture with multi-head self-attention fusion achieves state-of-the-art performance in detecting fake accounts, botnets, and coordinated inauthentic behaviour, surpassing classical and graph-based baselines.

Future work will extend the framework to support real-time stream processing for large-scale social media platforms, incorporate image and video content analysis for multimodal fake profile detection, and develop cross-platform detection capabilities. Explainability modules will be added to provide interpretable reasoning for each fake account flagged, supporting transparent moderation decisions.

. REFERENCES

[1] Y. Zhang et al., "BERT-based fake account detection on Twitter," IEEE ICSOC, 2021.

[2] S. Ahmed and M. Khan, "Bot detection using RoBERTa embeddings," ACM CCS, 2020.

[3] R. Kumar et al., "Transformer-GNN hybrid for fake account detection on Facebook," ACM WebSci, 2022.

[4] X. Li et al., "Coordinated fake account detection using BERT and behaviour sequences," AAAI, 2021.

[5] J. Wang and L. Chen, "Multi-modal fake profile detection using Vision-Language Transformers," WWW, 2023.

[6] S. Cresci et al., "Social fingerprinting: Detection of spambot groups through DNA-inspired behavioural modelling," IEEE TKDE, vol. 30, no. 8, 2019.

[7] A. Vaswani et al., "Attention is all you need," NeurIPS, 2017.