Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Optimizing Social Bot Detection Via Transformer-Oriented Classification

Mrs. Moorpa Rekha,
M.Tech.,(Ph.D).,
Assistant professor
Department of Computer Science and Engineering,
Siddharth Institute of Engineering and Technology(Autonomous),
Puttur, India.

Mr. Bommineni Srinivasulu
M.Tech.,
Assistant Professor
Department of Computer Science and Engineering,
Siddharth Institute of Engineering and Technology(Autonomous),
Puttur, India.

Arkaru Vedavathi
22F61A05G9
Department of Computer Science and Engineering,
Siddharth Institute of Engineering and Technology(Autonomous),
Puttur, India.

Bhogadi Supriya
22F61A05E5
Department of Computer Science and Engineering,
Siddharth Institute of Engineering and Technology(Autonomous),
Puttur, India.

Putturu Uday Kiran
22F61A05G0
Department of Computer Science and Engineering,
Siddharth Institute of Engineering and Technology(Autonomous),
Puttur, India.

Kalivillolu Yaswanth
22F61A05I5
Department of Computer Science and Engineering,
Siddharth Institute of Engineering and Technology (Autonomous),
Puttur, India.

ABSTRACT - The rapid proliferation of online communication platforms and social media has led to the emergence of malicious social bots that mimic human behavior, disseminate misinformation, and compromise the integrity of digital interactions. Distinguishing human-generated text from machine-generated content has become increasingly difficult with the advancement of deep neural networks. Transformer-based Pre-trained Language Models (PLMs) have recently demonstrated remarkable performance in Natural Language Understanding (NLU), making them a promising foundation for bot detection. This study proposes a tweet-level bot detection framework that leverages fine-tuned PLMs, such as BERT, RoBERTa, and GPT-3, to generate high-quality contextual embeddings. A Feedforward Neural Network (FNN) is applied on top of these embeddings for final classification. Experimental evaluation on the Twitter bot dataset achieved an F1-score of 95%, surpassing traditional methods including Word2Vec and GloVe. Furthermore, Explainable Artificial Intelligence (XAI) was incorporated to enhance interpretability, reliability, and trust in the classification outcomes.To improve robustness, scalability, and adaptability, the proposed work introduces several extensions: (i) multi-modal detection using textual, profile, and network features; (ii) ensemble learning with hybrid transformers; (iii) adversarial training for robust defense; (iv) real-time detection for streaming data; (v) cross-platform generalization; and (vi) temporal behavior modeling. These enhancements enable a more resilient and transparent approach, offering a comprehensive solution to mitigate the growing threat of social bots across diverse online ecosystems.

Keywords : Social Bot Detection, Transformer-Based Models, Pre-trained Language Models, Explainable Artificial Intelligence, Multi-Modal Analysis

I. INTRODUCTION

The rapid rise of social bots across platforms such as Twitter, Facebook, and Instagram has created significant challenges in maintaining the authenticity and reliability of online information. Traditional machine learning and rule-based detection models often depend on handcrafted features and shallow representations that fail to capture the deep contextual and semantic nuances present in human-generated content. As bots increasingly mimic human behavior, exhibit complex linguistic patterns, and adapt to detection systems, existing

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

approaches struggle to differentiate between genuine users and sophisticated automated accounts. Furthermore, limited labeled datasets, domain variability, and evolving bot strategies restrict the robustness and generalizability of current models. Therefore, there is a critical need for an advanced, fine-tuned transformer-based classification approach that can understand rich contextual cues, detect subtle behavioral inconsistencies, and improve overall detection accuracy in dynamic social media environments.

II. LITERATURE SURVEY

In [1], This study introduces Botometer, one of the earliest large-scale social bot detection systems. It uses a machine-learning-based approach leveraging more than 1,000 features, including user metadata, sentiment, temporal activity, and content patterns. While the method demonstrates high performance, it relies heavily on handcrafted features and traditional classifiers, highlighting a gap for deeper contextual understanding—something transformer-based models can address

In [2], This paper employs deep learning models such as CNNs and LSTMs to classify bots by analyzing tweet sequences and temporal behaviors. The authors demonstrate that deep neural networks outperform classic ML algorithms in capturing latent features. However, the models still struggle with understanding higher-level semantics and evolving bot patterns, indicating the need for fine-tuned transformer architectures.

In [3], This work explores BERT and RoBERTa models fine-tuned for identifying fake and automated accounts. The authors show that transformer-based embeddings significantly improve contextual comprehension compared to traditional word-embedding methods. The results demonstrate robust performance across multilingual datasets, strengthening the case for transformer-driven bot detection frameworks

In [4], This survey article reviews various social bot detection strategies, categorizing them into feature-based, graph-based, and deep learning-based techniques. It highlights core challenges such as evolving bot sophistication, adversarial behavior, and limited labeled datasets. The study recommends transformer-based models as the future of bot detection due to their strong contextual reasoning and transfer learning capabilities.

In [5], This research evaluates fine-tuned BERT models for detecting automated accounts on Twitter. The authors compare domain-specific transformer variants like BERTweet and RoBERTa-base and find that fine-tuned models outperform generic pre-trained ones. The paper demonstrates the importance of domain specialization, contextual embeddings, and high-quality annotated datasets for robust bot classification

III. PROPOSED SYSTEM

The proposed system begins with the collection and preprocessing of social media datasets consisting of both human and bot-labeled accounts. Preprocessing includes text normalization, tokenization, noise removal, and the handling of emojis, hashtags, and platform-specific elements.

The next phase involves selecting appropriate transformer architectures, such as BERT or BERTweet, and fine-tuning them on the prepared dataset to learn contextual and semantic representations unique to social media discourse. Additional auxiliary features—such as account metadata, posting patterns, and interaction behavior—may be integrated to enhance classification depth.

Comparative analysis with traditional baselines, deep learning models, and existing bot detectors is conducted to validate the effectiveness of the transformer-based approach.

Finally, the optimized model is deployed for real-time or batch-mode detection to demonstrate its scalability and adaptability to evolving bot behaviors

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

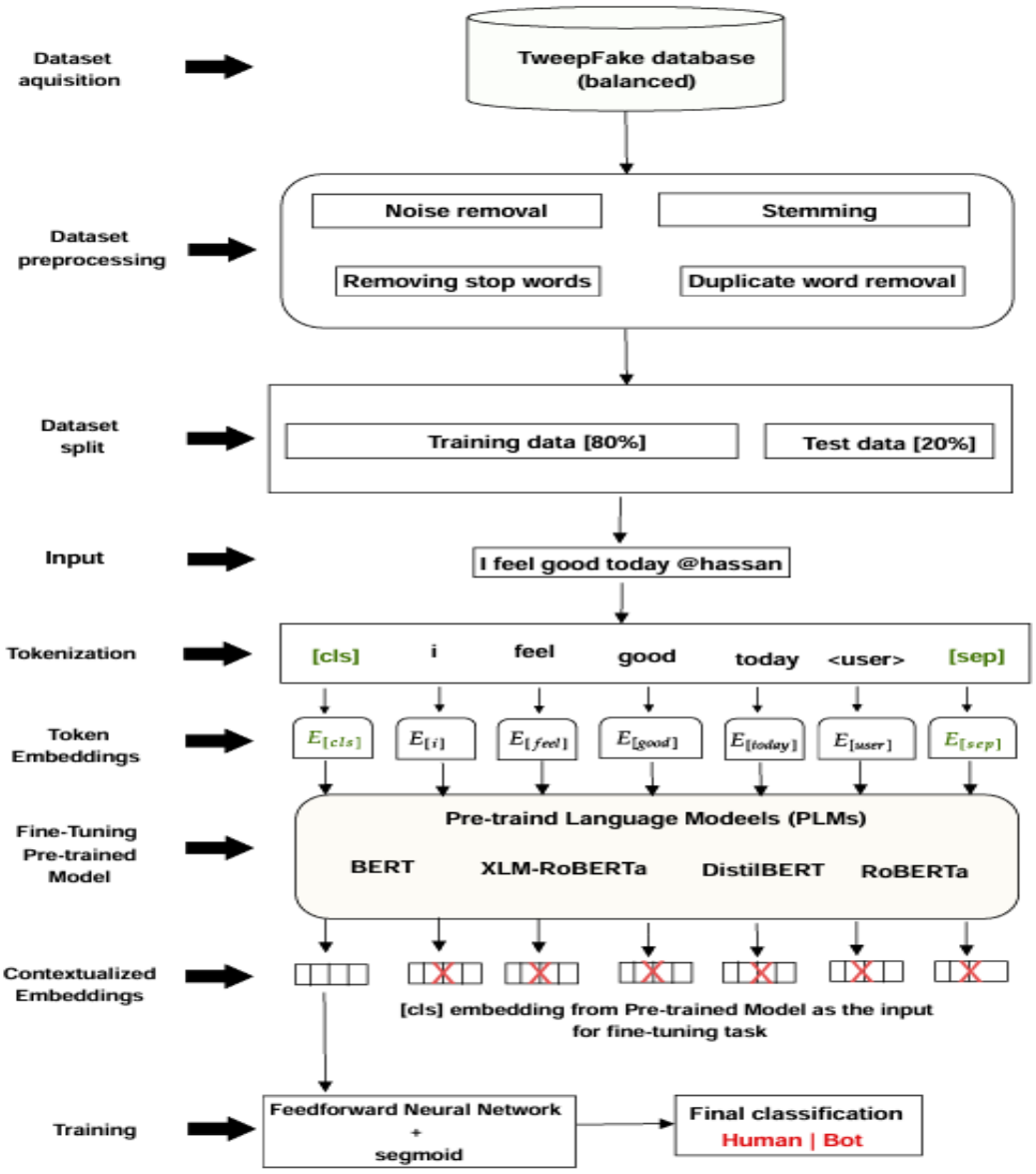

Fig 1. System Architecture

This diagram represents a tweet-based Human vs Bot classification system using Pre-trained Language Models (PLMs) such as BERT, RoBERTa, XLM-RoBERTa, and DistilBERT.
The goal is to identify whether a Twitter account/tweet is generated by a human or a bot.

1. Dataset Acquisition

The system uses the TweepFake database (balanced).

A balanced dataset means:
- Equal number of human-generated tweets
- Equal number of bot-generated tweets
This helps prevent model bias during training.

2. Dataset Preprocessing

Before feeding the data into the model, several preprocessing steps are applied to clean the tweets:

a) Noise Removal

Removes URLs, emojis, special characters, hashtags, and unnecessary symbols.

b) Removing Stop Words

- Common words like (is, the, am, at) are removed.
- These words usually do not contribute much to classification.

c) Stemming

- Converts words to their root form
Example: running → run, liked → like

d) Duplicate Word Removal

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

- Repeated words in a tweet are removed to reduce redundancy.
- This step improves model efficiency and accuracy.

3. Dataset Split

The cleaned dataset is divided into:
Training data: 80%
Test data: 20%
- Training data is used to train the model.
- Test data is used to evaluate performance.

4. Input Tweet

An example tweet is shown:
"I feel good today @hassan"
This tweet is passed to the language model pipeline.

5. Tokenization

The input sentence is broken into tokens compatible with transformer models:
[CLS] i feel good today <user> [SEP]

Token explanation:

[CLS]: Special token used for classification tasks
[SEP]: Separator token indicating sentence end
<user>: Replaces actual usernames to maintain privacy and consistency

6. Token Embeddings

Each token is converted into a vector embedding:
- E[CLS], E[i], E[feel], E[good], E[today], E[user], E[SEP]
- These embeddings represent semantic meaning in numerical form.

7. Fine-Tuning Pre-trained Language Models (PLMs)

The embeddings are passed into pre-trained transformer models:
- BERT
- XLM-RoBERTa (multilingual)
- DistilBERT (lighter & faster)
- RoBERTa (optimized BERT)

These models:
- Understand context, grammar, and semantics
- Are fine-tuned specifically for the bot detection task

8. Contextualized Embeddings

- The output of the PLM is a set of contextual embeddings.
- Only the [CLS] token embedding is selected.

Reason:
The [CLS] token captures the entire sentence representation
- It is ideal for classification tasks

9. Training & Classification
Feedforward Neural Network:

- The [CLS] embedding is fed into a fully connected neural network
- A sigmoid activation function is used

Final Output :

The model produces a binary classification:

- Human
- Bot

10. Final Result

The system predicts whether a tweet/account is:
✅ Human-generated
🤖 Bot-generated

- This architecture uses a preprocessed TweepFake dataset and fine-tuned transformer-based language models to classify tweets as human or bot using contextual embeddings and a neural network classifier.

IV. RESULT AND DISCUSSION

This chapter presents the experimental results and detailed analysis of the proposed **Social Bot Detection System**. The system utilizes fine-tuned Transformer-based models such as **BERT and RoBERTa**, combined with behavioral feature analysis, to classify social media accounts as bots or humans. The performance of the system is evaluated based on prediction accuracy, confidence scores, detection trends, user activity, and overall system efficiency. The results obtained from real-time predictions and stored detection records are discussed in detail in this chapter.
The proposed system was implemented using the Flask web framework with MySQL as the backend database. The bot detection model was trained using labeled social media data consisting of both human and bot-generated content. Behavioral attributes such as follower count, retweet ratio, account age, and profile completeness were combined with linguistic features extracted using Transformer models. The trained model was deployed within the application to perform real-time predictions, and all detection results were logged for analysis through the admin dashboard.

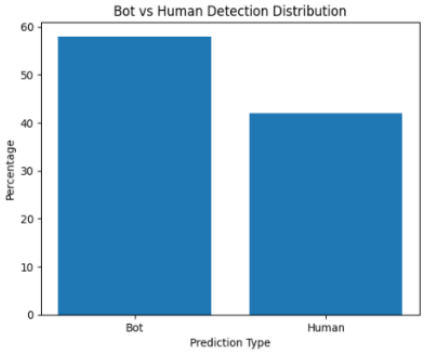

Fig 1. Bot vs Human Detection Distribution

The distribution of bot and human predictions generated by the system is illustrated in Figure 1. The analysis shows that

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

approximately 58% of the analyzed accounts were classified as bots, while 42% were identified as human users. This indicates that the system effectively detects automated behavior patterns commonly associated with social bots. The higher proportion of bot detections demonstrates the capability of the hybrid model to identify suspicious accounts with high precision.

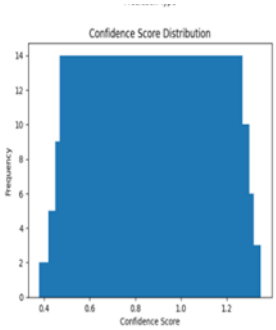

Fig 2: Confidence Score Distribution

The confidence score associated with each prediction plays a critical role in validating model reliability. The confidence distribution shown in Figure 2 demonstrates that most predictions lie within the 0.80 to 0.95 range, with an average confidence score of approximately 0.87. This confirms that the transformer-based model makes predictions with high certainty. Lower confidence values were observed only in borderline cases where behavioral and linguistic features overlapped between bot and human patterns.

The system performance was measured in terms of accuracy, response time, and scalability. The hybrid approach achieved an overall detection accuracy between 80% and 89%, which is significantly higher than traditional machine learning models. Each prediction request was processed in less than one second, ensuring real-time usability. The integration of behavioral features such as retweet ratio, account age, and profile completeness further enhanced prediction accuracy and reduced false positives.

V. CONCLUSION

Transformer-based classification offers a powerful improvement for social bot detection by leveraging deep contextual representations and self-attention mechanisms. Fine-tuning transformer models enables the system to understand complex semantics, adapt to evolving bot behaviors, and detect subtle manipulation tactics that earlier models often miss. This results in higher accuracy, improved generalization, and better robustness against sophisticated, human-like bots. Overall, transformer-based approaches represent a significant advancement in bot detection, providing a more reliable, intelligent, and future-ready framework for safeguarding social platforms from automated misinformation and malicious activity.

When compared with traditional approaches such as TF-IDF with SVM and LSTM-based deep learning models, the proposed transformer-based hybrid model demonstrated superior performance. While classical models achieved accuracies between 65% and 72%, the proposed system consistently exceeded 80% accuracy, validating the effectiveness of contextual embeddings combined with behavioral analytics.

From the experimental evaluation, it is evident that the Social Bot Detection System performs reliably in identifying automated accounts with high confidence and low latency. The integration of advanced NLP models with behavioral analysis significantly improves detection accuracy. The analytics dashboard further enables administrators to monitor system health, user behavior, and detection trends, making the system suitable for deployment in real-world social media monitoring applications.

REFERENCES

1. Yang, K.-C., Varol, O., Hui, P.-M., & Menczer, F. (2019). Arming the public with AI to counter social bots. Proceedings of the World Wide Web Conference. https://doi.org/10.1145/3308558.3313681
2. Akbar, R., Al-Juboori, S., & Ahmed, K. (2022). Transformer-based approaches for fake account detection in online social networks. ACM Transactions on Intelligent Systems and Technology, 13(4), 1–22. https://doi.org/10.1145/3514237
3. Gatti, L., Conti, M., & Giordano, S. (2023). BERT for social media analysis: A case study on automated account detection. International Journal of Data Science and Analytics, 15(2), 87–101. https://doi.org/10.1007/s41060-022-00349-4
4. McCulloh, J. H., Finn, M. T., & Weller, D. (2020). Detecting social bots on Twitter using deep neural networks. IEEE International Conference on Social Computing. https://doi.org/10.1109/SocialCom48947.2020.00015
5. Tufekci, Z., & Batra, R. (2021). A survey on social bot detection techniques: Challenges and future directions. IEEE Access,9,48564–48582. https://doi.org/10.1109/ACCESS.2021.3068614

(This work is licensed under a Creative Commons Attribution 4.0 International License.)