International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 03(March),2026

Received: 03-02-2026 Accepted: 10-03-2026 Published: 17-03-2026

TRUSTLENS: EVALUATING TWEET AUTHENTICITY THROUGH PROFILE AND CONTENT

Vandamsetty Bindu Madhuri [1], **Nali Srivani** [2], **Yadlapalli, Yadlapalli Bhuvana Priya** [3], **Veeramallu Hanumath Valli Sravan** [4], **Puppala Sowmya** [5]

[1] Dept. of CSE, SR Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India,
bindumadhurivandamsetty@gmail.com.

[2] Assistant Professor, Dept. of CSE, SR Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India.
[3] Dept. of CSE, SR Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India.
[4] Dept. of CSE, SR Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India.
[5] Dept. of CSE, SR Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India.

Abstract— *Social media platforms such as Twitter play a major role in information sharing; however, they are increasingly affected by fake tweets and spam accounts that spread misleading or harmful content. Detecting such content manually is impractical due to the large volume and real-time nature of social media data. This paper presents TrustLens, a hybrid approach for evaluating tweet authenticity by combining content-based deep learning analysis with user profile evaluation. A Long Short-Term Memory (LSTM) neural network is employed as the core component of the proposed system to analyze tweet textual content and classify it as spam or non-spam. In addition, user profile attributes such as account age, follower–following ratio, verification status, and activity patterns are analyzed to estimate the spam risk of an account. The final prediction is obtained by integrating both content-based and profile-based assessments using a weighted fusion strategy. Experimental results on a Twitter spam dataset demonstrate that the proposed approach achieves high accuracy and improves detection reliability compared to traditional machine learning models.*

Keywords— Twitter spam detection, tweet authenticity, machine learning, deep learning, LSTM, user profile analysis

I. INTRODUCTION

Online social networking platforms now play a central role in digital communication, information exchange, and the shaping of public discourse. Among these platforms, Twitter enables users to rapidly disseminate short textual messages, commonly known as tweets. However, the open nature of Twitter has also led to a significant rise in fake tweets, spam content, and malicious user accounts, which can mislead users and cause social, economic, and political harm. Detecting such content manually is challenging due to the massive volume and real-time nature of data generated on the platform.

Over the past few years, data-centric techniques leveraging machine learning and deep learning models have gained significant attention for solving a broad spectrum of classification and forecasting problems spanning different application domains. Deep neural architectures, particularly Convolutional Neural Networks (CNNs) and intelligent learning-based frameworks have demonstrated strong performance for learning intricate data representations that support accurate predictive outcomes. Optimization-based learning and feature selection techniques have also been explored to enhance model performance and robustness in classification systems. These findings demonstrate that advanced learning models are capable of effectively processing complex, high-dimensional data.

Several existing approaches for Twitter spam and fake content detection primarily focus on tweet textual analysis using traditional machine learning algorithms. While such methods achieve reasonable accuracy, they often ignore important user profile characteristics such as account age, follower–following ratio, and verification status. As a result, text-only models may fail to detect coordinated spam campaigns or newly created malicious accounts.

To overcome these limitations, this paper proposes TrustLens, a hybrid deep learning–based approach for evaluating tweet authenticity on Twitter. The proposed system combines an LSTM-based tweet content analysis model with user profile–based spam risk evaluation. Tweet content is processed using a Long Short-Term Memory (LSTM) network, whereas user-related metadata are evaluated through behavioral feature analysis. The final decision is obtained by integrating content-based and profile-based predictions using a weighted fusion strategy. The evaluation outcomes indicate that the proposed method achieves higher detection accuracy and improved reliability when compared with conventional machine learning techniques.

II. Objectives

This study focuses on designing a comprehensive framework that assesses tweet authenticity by detecting deceptive content and identifying spam-oriented user accounts on Twitter. The proposed system aims to leverage deep learning techniques to analyze tweet textual content and accurately classify it as spam or non-spam. By learning sequential patterns within tweet text, the system seeks to improve detection accuracy compared to traditional machine learning approaches.

Another important objective of this study is to incorporate user profile–based behavioral analysis into the spam detection process. User attributes such as account age, follower–following ratio, verification status, activity levels, and profile completeness are analyzed to estimate the spam risk associated with an account. This objective addresses the limitations of text-only detection systems by enabling the identification of newly created or behaviorally abnormal spam accounts.

Beyond the primary objective, the proposed framework also aims to integrate content-based tweet analysis and user profile evaluation using a hybrid decision fusion strategy. By combining predictions from the deep learning model and the profile-based spam risk assessment, the system seeks to enhance robustness and reduce false positives and false negatives. This integrated approach is designed to provide more reliable and consistent tweet authenticity evaluation.

In addition, the proposed system aims to support real-time tweet analysis by utilizing Twitter API data. The framework is designed to accept tweet URLs or custom text as input and generate interpretable outputs such as spam classification results, class probabilities, and confidence scores. Through these objectives, the proposed TrustLens framework seeks to improve the effectiveness, reliability, and practical applicability of social media spam detection systems.

III. Literature review

Learning-based computational techniques have gained widespread adoption across multiple application domains to address complex classification and prediction problems. Content-based learning approaches using neural networks have shown strong effectiveness in capturing and modeling user-specific patterns and feature representations. Markapudi *et al.* introduced a content recommendation approach based on neural network architectures that achieved improved prediction accuracy through effective feature extraction and learning mechanisms [1]. Similarly, machine learning models have been applied to customer behavior analysis, where comparative studies on predictive models have shown their effectiveness in classification tasks such as customer churn prediction [2]. Fusion-based learning approaches have also been successfully utilized in medical diagnosis, demonstrating improved classification performance by combining multiple learning strategies [3].

Recent advancements in deep learning have further strengthened the ability of models to extract meaningful patterns from complex, high-dimensional feature spaces. Convolutional Neural Networks (CNNs), especially, have demonstrated strong capability in learning complex representations for accurate prediction tasks. Guttikonda *et al.* employed CNN and machine learning techniques for accurate identification problems, highlighting the strengths of deep learning approaches in handling complex data representations [4].

Beyond deep learning architectures, researchers have investigated optimization-driven learning strategies to enhance prediction performance through effective feature selection and regularization. Techniques such as LASSO regularization combined with bio-inspired optimization algorithms have been shown to improve model robustness and accuracy. Guttikonda *et al.* reported that optimization-assisted learning models significantly enhance prediction outcomes in data-driven systems [5], [6].

Although these studies demonstrate the effectiveness of advanced learning models across diverse domains, their direct application to social media spam detection remains limited. Many existing Twitter spam detection approaches rely primarily on textual analysis and traditional machine learning algorithms, often ignoring user profile characteristics. This limitation motivates the development of a hybrid framework that integrates deep learning–based tweet content analysis with user profile–based behavioral evaluation for reliable tweet authenticity detection.

IV. Proposed Methodology

The proposed TrustLens framework is designed as a hybrid system that evaluates tweet authenticity by combining content-based deep learning analysis with user profile–based behavioral evaluation. The overall methodology consists of data collection, text preprocessing, tweet content classification using deep learning, user profile spam risk assessment, and final decision fusion. The workflow of the proposed system is illustrated through these stages.

A. Data Collection

Tweet data and corresponding user metadata are collected using the Twitter API. For experimental evaluation, a publicly available Twitter spam dataset containing labeled tweet text is used for the training and evaluation of the content-based classification model. The dataset consists of spam and non-spam tweets, enabling binary classification. For real-time analysis, tweet content and user profile information such as account age, follower count, following count, verification status, and activity metrics are retrieved through the Twitter API.

B. Text Preprocessing

Prior to classification, tweet textual content is processed through multiple preparation stages to convert unstructured text into a representation suitable for deep learning architectures. Tokenization is applied to convert words into numerical indices based on a predefined vocabulary. To ensure uniform input length across all samples, tokenized sequences are padded or truncated to a fixed length. Additionally, a vocabulary size limit is imposed to eliminate infrequent and noisy terms, thereby reducing computational complexity and improving model generalization. These preprocessing steps enable efficient learning of sequential and contextual patterns present in tweet text.

C. LSTM-Based Tweet Content Classification

The core of the content-based analysis is a Long Short-Term Memory (LSTM) neural architecture that effectively models sequential dependencies within text data and capturing long-range dependencies. An embedding layer is used to map input tokens into dense vector representations that preserve semantic relationships between words. This is followed by stacked LSTM layers that learn contextual dependencies within tweet sequences. Dropout layers are incorporated to mitigate overfitting and improve model robustness. The network’s final layer applies a sigmoid activation function to generate probabilistic output to estimate the probability of a tweet being spam or non-spam. Model training is carried out using a binary cross-entropy objective function, with optimization performed through the Adam algorithm to achieve reliable convergence.

D. User Profile Spam Risk Evaluation

In addition to tweet content analysis, user profile characteristics are evaluated to estimate the likelihood of an account being spam. Behavioral indicators such as follower–following ratio, account age, tweet frequency, verification status, and profile completeness are analyzed to compute a spam risk score. Accounts exhibiting abnormal behavioral patterns, such as newly created profiles with disproportionately high following counts or minimal engagement, are assigned higher spam risk values. Conversely, verified and well-established accounts with consistent activity patterns are assigned lower risk scores. This behavioral assessment provides complementary information that enhances the reliability of spam detection.

E. Hybrid Decision Fusion

The final tweet authenticity decision is obtained by integrating the outputs of the content-based LSTM classifier and the user profile spam risk evaluation. A weighted fusion strategy is employed, where greater emphasis is

Received: 03-02-2026 Accepted: 10-03-2026 Published: 17-03-2026

placed on the content-based prediction while the user profile risk contributes additional behavioral context. The combined probability score is compared against a predefined threshold to classify a tweet as spam or non-spam. This hybrid decision mechanism improves detection robustness by jointly leveraging textual semantics and user behavior, thereby reducing false positives and improving overall classification reliability.

Fig. 1 summarizes the overall operational flow of the proposed TrustLens framework, depicting the interaction between tweet content analysis and user profile–based evaluation.

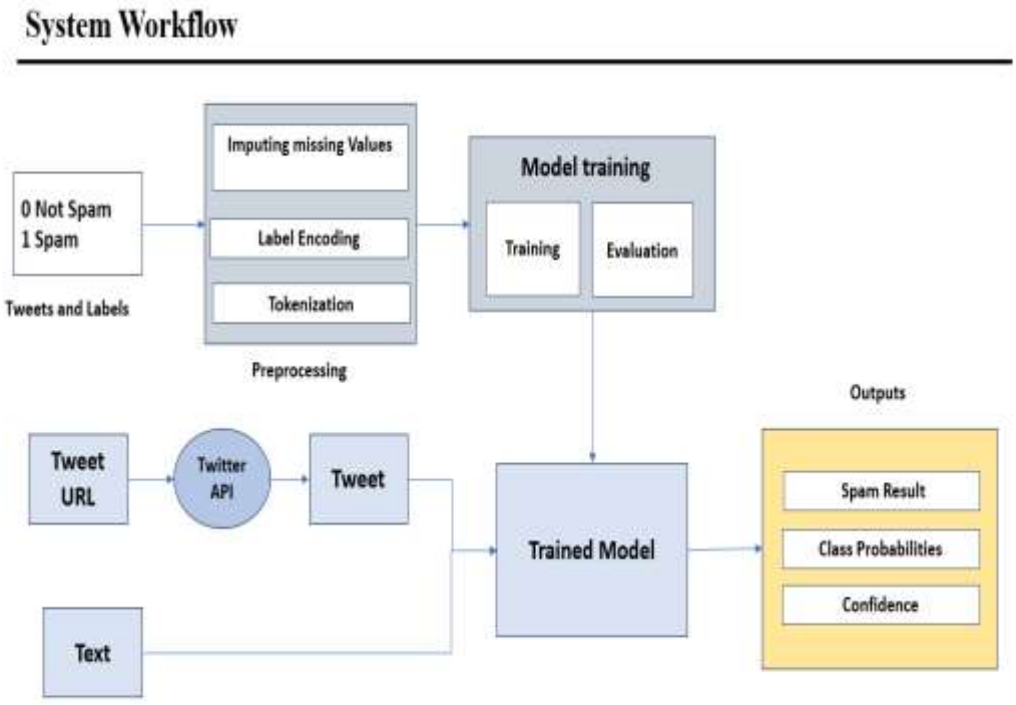


Fig. 1 operational flow of the proposed TrustLens framework

V. EXPERIMENTAL RESULTS AND DISCUSSION

This section describes the experimental assessment of the proposed TrustLens framework and discusses its effectiveness in detecting fake tweets and identifying spam accounts. The performance of the proposed approach is assessed using standard classification metrics, including accuracy, precision, recall, and F1-score, which collectively provide an in-depth evaluation of classification reliability and robustness.

A. Experimental Setup

The proposed model is developed and validated using a publicly available Twitter spam dataset consisting of labeled tweet text categorized as spam and non-spam. To ensure a fair evaluation, the dataset is partitioned into separate training and testing sets, where 80% of the samples are allocated for model training and the remaining 20% are used for evaluation. The LSTM-based deep learning model is trained for multiple using the Adam optimizer along with a binary cross-entropy loss formulation to achieve stable convergence.

For comparative evaluation, baseline machine learning classifiers, including Support Vector Machine (SVM), Decision Tree, Random Forest, and XGBoost, are implemented using TF-IDF–based textual feature representations. This experimental setup enables a direct comparison between deep learning–based sequential modeling and conventional text classification approaches.

B. Performance Evaluation

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 03(March),2026
Received: 03-02-2026 Accepted: 10-03-2026 Published: 17-03-2026

The effectiveness of the proposed LSTM-based model is assessed using standard performance measures, including accuracy, precision, recall, and F1-score. Results from the experiments demonstrate that the LSTM approach delivers better performance than traditional machine learning methods, particularly with respect to accuracy and F1-score. The observed improvement can be explained by the ability of the LSTM architecture to capture sequential dependencies and contextual information present in tweet text.

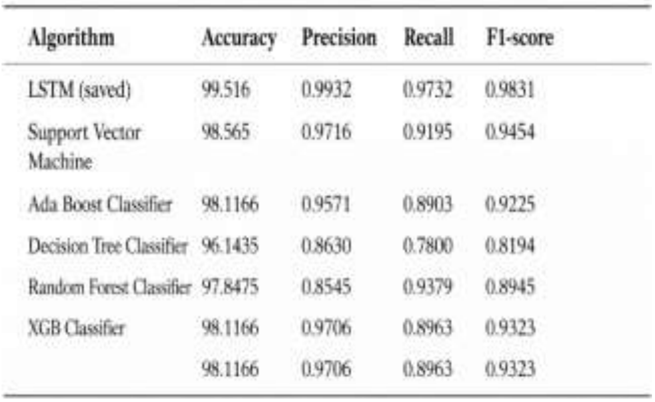
Furthermore, the integration of user profile–based spam risk evaluation enhances detection reliability by reducing misclassification cases. The combined analysis helps mitigate incorrect classifications, particularly in scenarios where textual content alone is insufficient to accurately identify spam behavior.

C. Comparative Analysis

A comparative analysis is conducted between the proposed TrustLens framework and baseline machine learning models. While traditional classifiers demonstrate reasonable performance for text-only spam detection, they are limited in their ability to capture behavioral characteristics associated with spam accounts. Models relying solely on textual features may fail to detect coordinated or newly created spam accounts with subtle content patterns.

In contrast, the proposed hybrid framework leverages both tweet content and user profile information, resulting in improved robustness and generalization capability. The experimental results confirm that integrating content-based deep learning with behavioral profile analysis leads to enhanced classification performance and more reliable spam detection on social media platforms.

Table I presents the comparative performance analysis of the TrustLens model with baseline machine learning classifiers.



Algorithm	Accuracy	Precision	Recall	F1-score
LSTM (saved)	99.516	0.9932	0.9732	0.9831
Support Vector Machine	98.565	0.9716	0.9195	0.9454
Ada Boost Classifier	98.1166	0.9571	0.8903	0.9225
Decision Tree Classifier	96.1435	0.8630	0.7800	0.8194
Random Forest Classifier	97.8475	0.8545	0.9379	0.8945
XGB Classifier	98.1166	0.9706	0.8963	0.9323
	98.1166	0.9706	0.8963	0.9323

Table. 2 comparative performance of the LSTM-based model

Table I illustrates the comparative performance of the proposed TrustLens LSTM-based model against traditional machine learning classifiers. The proposed approach achieves the highest accuracy and F1-score, indicating superior tweet spam detection capability. The results demonstrate that integrating deep learning-based content analysis with user profile–level evaluation improves detection reliability compared to conventional machine learning methods

VI. CONCLUSIONS

This paper presented TrustLens, a hybrid framework for evaluating tweet authenticity by integrating content-based deep learning analysis with user profile–level behavioral assessment. An LSTM-based model was employed to analyze tweet textual content, while user attributes such as account age, follower–following patterns, and activity metrics were utilized to estimate spam risk. The final prediction was obtained through a weighted fusion of content and profile-based evaluations. Experimental results on a publicly available Twitter spam dataset demonstrated that the proposed approach achieves superior classification effectiveness measured using accuracy and F1-score when compared with traditional machine learning classifiers. The findings confirm that combining tweet content analysis with user profile characteristics significantly enhances the reliability of fake tweet and spam account detection on social media platforms.

Received: 03-02-2026 Accepted: 10-03-2026 Published: 17-03-2026

ACKNOWLEDGMENT

The authors would like to express their gratitude to the management and administration of SR Gudlavalleru Engineering College for providing the necessary infrastructure and facilities to carry out this research. We also thank the Department of Computer Science and Engineering for their constant encouragement and technical support throughout the duration of this study.

REFERENCES

[1] B. Markapudi, K. Chaduvula, D. N. V. S. L. S. Indira, and M. V. N. S. S. R. K. S. Somayajulu, “Content-based video recommendation system (CBVRS): A novel approach to predict videos using multilayer feed forward neural network and Monte Carlo sampling method,” Multimedia Tools and Applications, vol. 82, no. 2, pp. 6965–6991, Aug. 2022

[2] K. J. Latha, M. Baburao, and K. Chaduvula, “A comparative study on logit leaf model and support leaf model for predicting customer churn,” International Journal of Computer Sciences and Engineering, vol. 7, no. 5, pp. 1628–1632, May 2019.

[3] M. Edupuganti, V. Rathikarani, and K. Chaduvula, “Classification of heart diseases using fusion-based learning approach,” International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 8s, pp. 570–580, Dec. 2023.

[4] K. Guttikonda, Y. Ashvitha, V. S. R. Reddy, R. M. Krishna, and P. Sandeep, “Integrating convolutional neural networks and machine learning for accurate identification of autism spectrum disorder using facial biomarkers,” in Proc. IEEE International Conference on Emerging Systems and Intelligent Computing (ESIC), Feb. 2024.

[5] K. Guttikonda, G. Ramachandran, and G. V. S. N. R. V. Prasad, “Autism spectrum disorder prediction using LASSO regularised bat search optimisation,” International Journal of Services Operations and Informatics, 2024.

[6] K. Guttikonda, G. Ramachandran, and G. V. S. N. R. V. Prasad, “Cuckoo search optimization-based feature selection for predicting autism spectrum disorder using artificial immune algorithms,” Journal of Theoretical and Applied Information Technology, Jan. 2025.

[7] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, “Detecting spammers on Twitter,” in Proc. 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS), 2010.

[8] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.

[9] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] C. Castillo, M. Mendoza, and B. Poblete, “Information credibility on Twitter,” in Proc. 20th International World Wide Web Conference (WWW), 2011.