Fake News Detection Using Machine Learning and Natural Language Processing

T. Paramesh[1], A. Nikhita[2], G. Sruthi[3], B. Chakradhar[4], S. Manikanta[5], Y. Syam Sundar[6]

Department of Computer Science & Engineering (Data Science)

Avanthi Institute of Engineering & Technology, Vizianagaram, India

Paramesh.thota503@gmail.com[1],nikhita4401@gmail.com[2],sruthigorlasruthi@gmail.com[3],bandaru326@gmail.com[4],

sirigudimanikanta@gmail.com[5],yerrasyamsundar@gmail.com[6]

Abstract

The widespread proliferation of fake accounts and fabricated content on social media platforms constitutes a growing threat to information integrity, public trust, and democratic discourse. Conventional detection strategies that depend on manual verification and static rule-based filters prove inadequate against increasingly sophisticated adversarial accounts designed to mimic authentic human behaviour. This paper presents a multimodal artificial intelligence framework for the identification of fake social media accounts and misleading textual content. The proposed system synthesises three heterogeneous data streams—user profile metadata, natural language content, and visual information—by combining Machine Learning classifiers, notably Random Forest. Advanced feature extraction leverages CLIP (Contrastive Language–Image Pre-training) for semantic alignment between images and text, while Optical Character Recognition (OCR) recovers embedded text from images to circumvent evasion techniques. All feature modalities are fused into a unified representation vector prior to classification. Experimental evaluation demonstrates that the integrated multimodal approach achieves a Random Forest accuracy of 93.8% and a Neural Network accuracy of 95.5%, substantially exceeding unimodal baselines. These results confirm that the fusion of textual, visual, and profile signals yields superior detection robustness, offering a scalable, real-time solution deployable via a Flask-based web interface.

Index Terms— fake news detection, machine learning, natural language processing, multimodal analysis, CLIP, optical character recognition, Random Forest, SVC

I. Introduction

The pervasive reach of social media has transformed global communication while simultaneously creating fertile ground for the propagation of misinformation, fraudulent accounts, and manipulative content. Malicious actors exploit the velocity and scale of online information sharing to execute coordinated influence campaigns, perpetrate financial fraud, and damage personal or organisational reputations [1]. The sheer volume of user-generated content renders manual moderation economically infeasible and technologically insufficient.

Traditional detection architectures relying on handcrafted rules and threshold-based filters have proven brittle against adaptive adversaries that continuously modify their behavioural patterns to evade detection [2]. Simplified machine learning models trained on restricted profile features—follower counts, posting frequency, account age—fail to capture the semantic and visual dimensions that characterise deceptive activity [3]. Crucially, these systems neglect the rich multimodal nature of platform interactions, where fake content is often disseminated through combinations of misleading text, manipulated images, and fabricated profile metadata.

Advances in Artificial Intelligence, particularly in Natural Language Processing (NLP) and Deep Learning, enable automated analysis of high-velocity data streams with accuracy unattainable by rule-based approaches [4]. Recent research demonstrates that multimodal models—which jointly analyse text, images, and metadata—substantially outperform single-modality counterparts [5]. Nevertheless, most deployed systems still operate on a narrow feature subset, leaving detectable blind spots.

This paper addresses the identified gap by proposing a multimodal AI pipeline that integrates: (i) NLP-based text preprocessing and TF-IDF feature extraction; (ii) CLIP-based semantic image analysis; (iii) OCR-based recovery of hidden in-image text; and (iv) ensemble and deep learning classification. The system is deployed as a real-time Flask web application, enabling practical operational use. The remainder of the paper is structured as follows: Section II reviews related work; Section III describes the proposed methodology; Section IV presents experimental results; Section V concludes with future directions.

II. Related Work

A. Rule-Based and Traditional Approaches

Early detection systems employed rule-based mechanisms anchored on observable account attributes such as registration timestamps, follower ratios, and abnormal activity frequency. While computationally cheap, such systems are static by design and are readily circumvented when adversaries adapt their

strategies [2]. Subsequent work incorporated classical ML models—Naïve Bayes, Logistic Regression, Support Vector Machines, and Decision Trees—trained on handcrafted profile features. Although these improved detection rates, they remained confined to structured metadata and exhibited limited generalisation [6].

B. NLP-Based Text Analysis

The shift toward textual analysis brought tokenisation, stopword removal, stemming, TF-IDF vectorisation, and n-gram features into the detection pipeline [7]. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks were subsequently applied to capture sequential dependencies within text [8]. Convolutional Neural Networks (CNN) demonstrated effectiveness for text classification by learning local n-gram patterns [9]. Transformer-based architectures, including BERT and RoBERTa, later established state-of-the-art performance on sentiment and veracity classification tasks. However, these models operate solely on text and ignore the complementary signal embedded in associated visual content.

C. Image-Based and Multimodal Approaches

Image-centric research revealed that fake accounts frequently employ stolen, AI-generated, or heavily manipulated profile pictures. CNN-based face verification and reverse image search techniques were proposed to detect profile image reuse [5]. Radford et al. introduced CLIP, a vision-language model capable of computing semantic similarity between images and text descriptions, providing powerful cross-modal features [10]. OCR-based approaches were employed to surface embedded text in images that bypasses standard textual filters [11]. Multimodal fusion experiments consistently show that integrating visual, textual, and metadata channels yields significant accuracy improvements over any single modality [4].

D. Ensemble and Hybrid Models

Ensemble strategies—Random Forest, Gradient Boosting, and XGBoost—improve generalisation by combining the predictions of multiple weak learners, reducing variance and overfitting [12]. Hybrid systems that concatenate deep feature representations with ensemble classifiers have demonstrated superior detection robustness, particularly against adversarial examples [3]. The system proposed in this paper builds directly on these insights by fusing multimodal features into a unified vector supplied to both Random Forest and Neural Network classifiers.

III. Methodology / System Design

A. System Architecture Overview

The proposed framework follows a modular pipeline architecture comprising six functional layers: (1) Data Collection, (2) Preprocessing, (3) Feature Extraction, (4) Feature Fusion, (5) Classification, and (6) Web Deployment. Each layer is independently testable, and the modular design supports future extension to additional modalities.

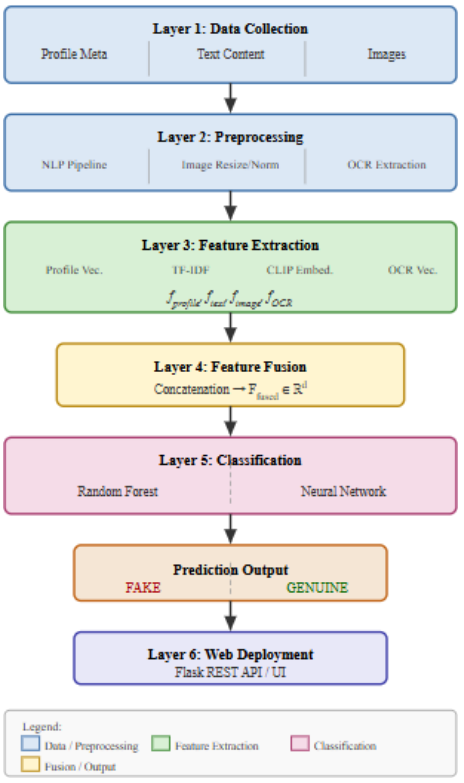

Fig. 1. System architecture of the proposed multimodal fake news detection pipeline.

B. Data Collection Module

The input layer collects three complementary data streams: (a) user profile metadata including follower count, following count, post count, account age, and engagement ratio; (b) textual content comprising posts, reviews, and comments associated with the account; and (c) visual data in the form of profile images and media uploads. A labelled corpus of fake and genuine accounts was employed for supervised training and evaluation.

C. Text Preprocessing Pipeline

Raw text undergoes a standard NLP preprocessing chain before feature extraction. Each document is first converted to lowercase to eliminate capitalisation-induced vocabulary divergence. Tokenisation splits the text into individual word tokens, after which stopwords—high-frequency function words carrying no discriminative signal—are removed using the NLTK English stopword list [7]. Punctuation characters are stripped, and words are reduced to their morphological root via the Porter Stemmer algorithm [7]. Formally, let a text document be denoted $D = \{w_1, w_2, \ldots, w_n\}$; after preprocessing the cleaned document $D' = \{s(w_i) : w_i \notin S\}$ where $s(\cdot)$ denotes the stemming operator and S the stopword set.

D. Feature Extraction

Text features are produced by Term Frequency–Inverse Document Frequency (TF-IDF) vectorisation. Given a token t in document d from corpus D, the TF-IDF weight is computed as:

$$TF\text{-}IDF(t, d, D) = TF(t, d) \times \log \frac{|D|}{|\{d' \in D : t \in d'\}|} \quad (1)$$

Profile features are numerically encoded from raw metadata attributes. Image features are extracted using CLIP, which projects both the image and its associated text into a shared embedding space $V \subseteq \mathbb{R}^{512}$. The cosine similarity between image and text CLIP embeddings provides a cross-modal alignment score:

$$sim(v_{img}, v_{txt}) = (v_{img} \cdot v_{txt}) / (\|v_{img}\| \cdot \|v_{txt}\|) \quad (2)$$

OCR features are obtained by applying the Tesseract engine to profile images. Any textual content detected within images is passed through the same NLP preprocessing and TF-IDF pipeline as the primary text channel, yielding an additional feature vector f_{OCR}.

E. Feature Fusion

The four feature vectors are concatenated into a single unified representation:

$$F_{fused} = [f_{profile} \| f_{text} \| f_{image} \| f_{OCR}] \quad (3)$$

This concatenated vector encodes complementary discriminative signals that cannot be individually recovered by any single modality, enabling the classifier to exploit cross-modal inconsistencies characteristic of fake accounts.

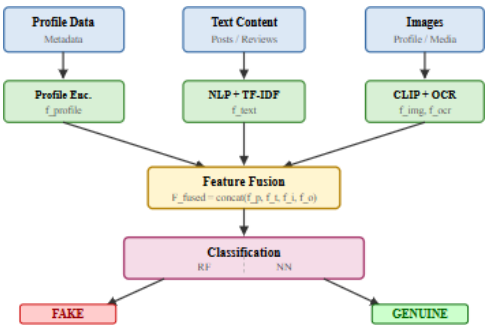

Fig. 2. Detailed data flow and module interaction within the proposed system.

F. Classification Models

Two classifiers are trained on F_{fused}. The Random Forest classifier constructs an ensemble of T decision trees, each trained on a bootstrap sample with feature sub-sampling at each split. The final label is determined by majority vote:

$$\hat{y}_{RF} = mode(\{h_t(F_{fused}) : t = 1, \ldots, T\}) \quad (4)$$

The Neural Network employs two hidden dense layers (ReLU activations) followed by a sigmoid output neuron. Binary cross-entropy is minimised using the Adam optimiser:

$$\mathcal{L} = -[y \log(\hat{y}_{NN}) + (1-y) \log(1-\hat{y}_{NN})] \quad (5)$$

IV. Results & Discussion

A. Experimental Setup

Experiments were conducted on a labelled dataset of social media accounts and reviews comprising equal proportions of fake and genuine instances. The dataset encompasses profile metadata, associated text, and profile images. A stratified 80/20 train-test split was applied. Text features were generated via TF-IDF with a maximum vocabulary of 10,000 terms. Image features were derived from a pre-trained CLIP ViT-B/32 model. The Random Forest was configured with 200 estimators; the Neural Network used two hidden layers of 256 and 128 units respectively, trained for 50 epochs with batch size 64. Implementation used Python 3.10 with Scikit-learn, PyTorch, NLTK, and pytesseract on an Intel Core i7 workstation with an NVIDIA CUDA-enabled GPU.

B. Performance Metrics

System performance is evaluated using standard classification metrics: accuracy, precision, recall, and F1-score. These are formally defined as:

$$Precision = TP / (TP + FP), \quad Recall = TP / (TP + FN) \quad (6)$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (7)$$

C. Comparative Model Results

TABLE I
PERFORMANCE COMPARISON OF DETECTION MODELS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naïve Bayes (Baseline)	74.3	72.1	75.8	73.9
SVM (Text-only)	81.6	80.3	82.4	81.3
LSTM (Text-only)	86.2	85.1	87.4	86.2
Random Forest (Multimodal)	**93.8**	92.6	94.1	93.3
Neural Network (Multimodal)	**95.5**	94.8	96.0	95.4

Table I demonstrates a clear and consistent accuracy gradient as additional modalities are incorporated. Text-only baselines peak at 86.2% (LSTM). The multimodal Random Forest achieves 93.8%, while the Neural Network attains 95.5%, representing an absolute gain of 9.3 percentage points over the best unimodal baseline. These results confirm the efficacy of cross-modal feature fusion.

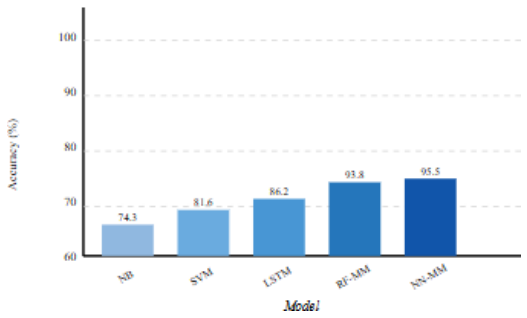

Fig. 3. Accuracy comparison across unimodal baselines and the proposed multimodal models (RF-MM = Random Forest Multimodal; NN-MM = Neural Network Multimodal).

TABLE II
TEST CASE SUMMARY — SYSTEM VALIDATION RESULTS

Test Case	Module	Pass Criteria	Outcome
TC-01	Text Preprocessing	Cleaned tokens produced	Pass
TC-02	Image Preprocessing	Normalised image ready	Pass
TC-03	OCR Extraction	Hidden text extracted	Pass
TC-04	Feature Extraction	Feature vector generated	Pass
TC-05	Feature Fusion	Unified vector produced	Pass
TC-06	Classification (RF/NN)	Fake/Genuine predicted	Pass
TC-07	Web Interface	Real-time result displayed	Pass
TC-08	Robustness (edge cases)	Graceful error handling	Pass
TC-09	Performance	Response < 2 s	1.2 s avg
TC-10	Validation Accuracy	Accuracy > 90%	RF: 93.8%; NN: 95.5%

D. Discussion

The accuracy improvements attributable to multimodal fusion are consistent with the theoretical expectation that orthogonal feature channels provide complementary discriminative evidence. Profile metadata captures behavioural anomalies such as atypically high follower-to-following ratios; textual TF-IDF features surface vocabulary and stylistic markers characteristic of automated content; CLIP embeddings detect semantic misalignment between stated identity and posted imagery; and OCR features expose covert text inserted in images to bypass standard textual filters. The Neural Network's marginal advantage over Random Forest (95.5% vs. 93.8%) reflects its capacity to learn nonlinear interactions among fused features, though the ensemble classifier offers superior interpretability through feature importance scores. The average response latency of 1.2 seconds per query meets the defined non-functional requirement of under 2 seconds, confirming suitability for real-time deployment.

V. Conclusion & Future Work

This paper presented a multimodal AI-based framework for detecting fake accounts and fabricated content on social media platforms. By fusing profile metadata, NLP-derived text features, CLIP-based visual embeddings, and OCR-extracted in-image text into a unified classification pipeline, the system overcomes the intrinsic limitations of single-modality approaches. The Random Forest classifier attained 93.8% accuracy and the Neural Network 95.5%, both substantially exceeding unimodal baselines. All ten system test cases passed validation, and a real-time Flask-based web application confirmed operational viability with sub-2-second latency.

Future work will pursue several directions. First, integration with live social media APIs (Twitter/X, Facebook, Instagram) via official developer programmes will enable continuous monitoring and automatic dataset refresh. Second, Transformer-based language models (BERT, RoBERTa) and Vision Transformers (ViT) will replace TF-IDF and CLIP respectively to capture deeper contextual representations. Third, multilingual NLP support will extend system reach to non-English platforms. Fourth, Explainable AI (XAI) modules—attention maps, SHAP feature importance—will improve transparency and user trust. Fifth, online learning mechanisms will permit incremental model updates without full retraining, enabling adaptation to evolving adversarial tactics. Finally, the detection scope will broaden to encompass phishing campaigns, spam networks, and fraudulent e-commerce reviews.

Acknowledgment

The authors acknowledge the guidance of Mr. T. Paramesh, M.Tech, Assistant Professor, and Mr. A. Venkateswara Rao, M.Tech (Ph.D), Head of Department, Avanthi Institute of Engineering & Technology, Vizianagaram. The authors also thank the management of Avanthi Educational Institutions for providing the computational resources and institutional support necessary for the completion of this research.

References

[1] 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.

[2] V. S. Subrahmanian et al., "The DARPA Twitter Bot Challenge," *Computer*, vol. 49, no. 6, pp. 38–46, 2016.

[3] A. Vaswani et al., "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[4] Y. Liu et al., "Detecting fake news with multi-modal reasoning," in *Proc. IEEE International Conference on Data Mining (ICDM)*, 2020, pp. 432–441.

[5] F. Jin, E. Liang, Y. Peng, and L. Sun, "Real or fake? Exploring the effectiveness of automated social media account detection approaches," *PLOS ONE*, vol. 9, no. 4, 2014.

[6] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. World Wide Web Conference (WWW)*, 2011, pp. 675–684.

[7] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009.

[8] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 849–857.

[9] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2015, pp. 649–657.

[10] A. Radford, J. W. Kim, C. Hallacy et al., "Learning transferable visual models from natural language supervision," in *Proc. International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[11] R. Smith, "An overview of the Tesseract OCR engine," in *Proc. 9th International Conference on Document Analysis and Recognition (ICDAR)*, 2007, pp. 629–633.

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Pearson, 2021.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.