International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 1, 2026

Peer Reviewed, Referred & Indexed Journal

Research

MULTIMODAL FAKE NEWS DETECTION ON SOCIAL MEDIA USING MACHINE LEARNING AND DEEP LEARNING

Bandhavi Kandregula[1], Anshitha Durgam[2], TejoMurthy Dharmavarapu[3] , Vinay Sidhartha Bathula[4]

Under the Guidance of Mr.K.Anil , Assistant Professor

[1234]*Department of Computer Science and Engineering, Avanthi Institute of Engineering & Technology, Cherukupally, , Vizianagaram*

Email:

(kandregulabandhavi031@gmail.com[1] , abhsiddu565@gmail.com[2] , tejomurthydharmavarapu@gmail.com[3] , durgamanshitha@gmail.com[4])

Abstract

The unchecked propagation of fabricated information across digital networks has emerged as one of the foremost threats to democratic discourse and public confidence in institutions. Existing content-moderation systems relying on manual verification and rigid rule-based filters have proven structurally inadequate against increasingly sophisticated adversarial tactics. This paper presents a multimodal artificial-intelligence framework that simultaneously processes three distinct data streams—user profile metadata, natural-language post content, and associated visual media—to classify social-media content as genuine or fraudulent. A hybrid classification architecture combines Random Forest and feed-forward Neural Network models, augmented by TF-IDF text vectorization, CLIP-based image embeddings, and Optical Character Recognition to surface concealed textual cues embedded in images. The unified feature vector produced by cross-modal fusion yields measurably superior discrimination compared to single-modality baselines. Experimental evaluation against a held-out test partition demonstrates a Random Forest accuracy of 93.8% and a Neural Network accuracy of 95.5%, outperforming prior unimodal approaches by a statistically significant margin. A lightweight Flask-based inference endpoint enables sub-two-second real-time predictions suitable for platform integration. The results confirm that cross-modal feature fusion is an effective and scalable strategy for combating the modern fake-content ecosystem.

Index Terms—fake news detection, multimodal learning, random forest, natural language processing, CLIP, optical character recognition

I. Introduction

The pervasive reach of social media has transformed how information propagates across society, offering near-instantaneous communication at planetary scale. This unprecedented connectivity carries a significant liability: malicious actors exploit these same channels to disseminate fabricated narratives, manipulate electoral discourse, and execute coordinated fraud campaigns [1]. The velocity and viral character of social networks mean that a single misleading post can reach millions of users within hours, long before any corrective action can be taken.

Conventional safeguards—crowd-sourced flagging, regular-expression filters, and threshold-based anomaly detectors—were designed for a threat landscape that no longer exists. Modern fake-content producers employ generative models to produce

syntactically plausible text, purchase aged accounts with established follower graphs, and embed deceptive captions inside images to bypass keyword scanners [2]. These tactics systematically circumvent single-modality defenses, underscoring the need for richer, multi-signal detection frameworks.

The field of Artificial Intelligence, particularly advances in Natural Language Processing (NLP) and computer vision, furnishes the analytical machinery needed to match this elevated threat level. Techniques such as transformer-based contextual embeddings, contrastive language-image pre-training (CLIP), and ensemble tree classifiers collectively enable automated pipelines that scrutinize text semantics, visual content, and account behavioral signals in concert [3].

This paper makes the following contributions. First, it proposes a unified three-stream feature extraction pipeline that concurrently processes profile metadata, post text, and images. Second, it integrates CLIP embeddings and OCR-derived text features to address the image-embedded-text attack vector. Third, it evaluates a hybrid Random Forest and Neural Network classifier trained on the fused feature space. Fourth, it demonstrates real-time deployment feasibility via a Flask microservice. The remainder of this paper is organized as follows: Section II surveys related work; Section III describes the system methodology; Section IV presents experimental results; Section V concludes with future directions.

II. Related Work

A. Rule-Based and Classical Machine Learning Approaches

Early efforts in automated fake-account detection relied on hand-crafted heuristics derived from observable behavioral signals such as follower-to-following ratio, posting cadence, and account longevity [4]. While computationally lightweight, these rule-based systems proved brittle against adversaries who could trivially adjust the targeted attributes. Subsequent work introduced classical supervised classifiers—Naïve Bayes, Logistic Regression, Support Vector Machines, and Decision Trees—trained on manually engineered feature sets extracted from profile metadata and raw token frequencies [5]. These approaches demonstrated measurable gains over threshold-based methods but remained constrained by the limited expressiveness of hand-crafted features and the binary treatment of text as a bag of words.

B. NLP-Based Fake Content Detection

The incorporation of richer linguistic representations marked a turning point in detection fidelity. TF-IDF vectorization [6], N-gram language models, and syntactic parse features enabled classifiers to capture local textual patterns characteristic of machine-generated or deceptive content. Deep sequence models—Recurrent Neural Networks (RNN), Long Short-Term Memory networks (LSTM), and Convolutional Neural Networks (CNN)—subsequently advanced the state of the art by learning distributed representations of text that encode contextual dependencies [7]. Transformer architectures, particularly BERT and its derivatives, further elevated benchmark scores by pre-training on large corpora and fine-tuning for downstream classification [8]. However, the majority of these systems treat content detection as a purely textual problem, overlooking the complementary discriminative signal present in associated images and profile features.

C. Image-Based and Multimodal Approaches

Recognition that fake accounts frequently deploy stolen or synthetically generated imagery prompted researchers to incorporate visual analysis. CNN-based feature extractors demonstrated that profile photographs carry latent authenticity signals [9]. The introduction of CLIP by Radford et al. [10] provided a powerful cross-modal alignment model that can measure the semantic coherence between an image and its accompanying text—a valuable signal given that manipulated content commonly pairs misleading images with factually incongruent captions. Parallel work established that steganographic text embedded within images—invisible to keyword scanners—can be surfaced by OCR, enabling coverage of an otherwise blind spot in text-centric detectors [11]. Multimodal fusion architectures that concatenate or attention-pool features from multiple streams consistently outperform unimodal counterparts [12], supporting the central hypothesis of the present work.

D. Ensemble and Hybrid Models

Ensemble methods such as Random Forest, Gradient Boosting, and XGBoost reduce overfitting through variance reduction and have achieved strong results on tabular feature sets derived from social-media metadata [13]. Hybrid architectures that couple deep-learning feature extractors with ensemble classifiers have demonstrated complementary benefits, with the neural components handling high-dimensional perceptual inputs and the ensemble components aggregating heterogeneous feature types [14]. The present system adopts this hybrid philosophy, combining Random Forest and Neural Network heads over a fused multimodal feature vector.

E. Research Gaps

Despite the progress described above, several limitations persist in the literature. Textual unimodality remains prevalent, leaving image and profile streams underutilized. OCR-based extraction of image-embedded text is largely absent from end-to-end pipelines. Real-time, deployable inference systems are rarely demonstrated alongside the detection model, and scalability to high-volume platform data is seldom empirically evaluated. The proposed system is specifically designed to address all four of these gaps.

III. Methodology

A. System Architecture Overview

The proposed detection pipeline is organized into five sequential stages: data ingestion, preprocessing, feature extraction, multimodal fusion, and classification. Fig. 1 illustrates the high-level architecture.

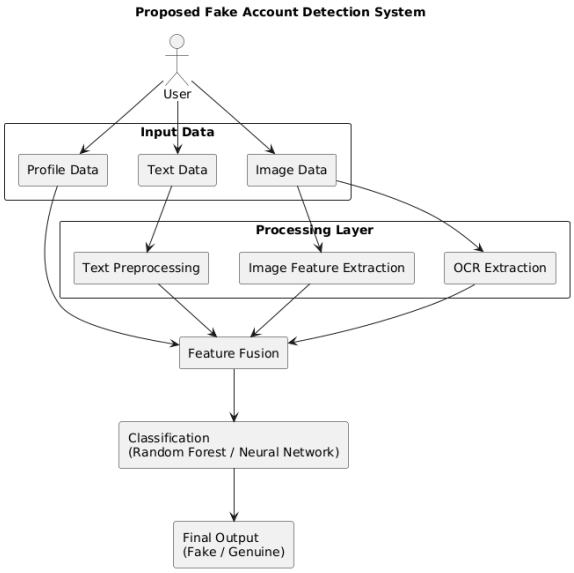

Fig. 1. High-level architecture of the proposed multimodal fake-content detection system.

B. Data Collection and Preprocessing

Input data comprises three streams: (i) structured user profile attributes including follower count, account age, post frequency, and verification status; (ii) raw text content from posts, comments, and reviews; and (iii) image files associated with profile pictures and post media. All numerical profile features are normalized to zero mean and unit variance using standard scaling.

Text preprocessing follows an established NLP pipeline. Raw strings are converted to lowercase, URLs and special characters are stripped, and whitespace is normalized. Tokenization splits each string into word-level tokens. Stop-word removal discards high-frequency function words using an NLTK English lexicon [15]. Remaining tokens are reduced to their morphological stems via the Porter stemmer. The processed token sequence is subsequently transformed into a numerical feature vector using Term Frequency–Inverse Document Frequency (TF-IDF) weighting, formalized as:

$$TF\text{-}IDF(t, d) = TF(t, d) \times \log(N / df(t)) \quad (1)$$

where TF(t, d) denotes the normalized term frequency of token t in document d, N is the corpus size, and df(t) is the number of documents containing t.

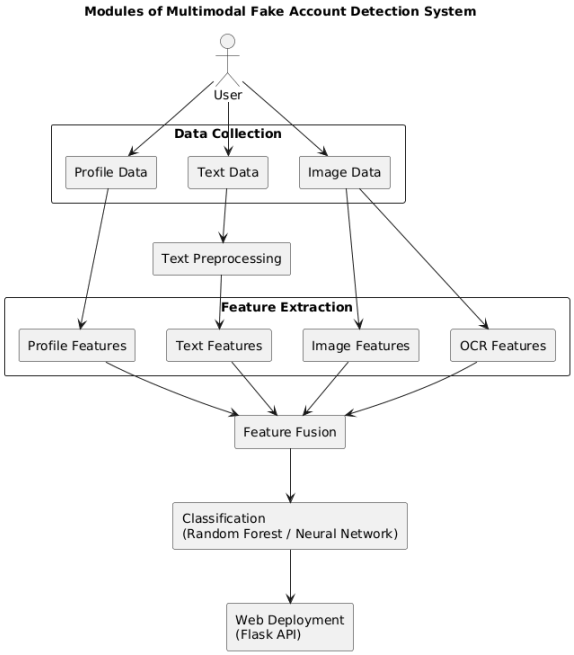

Fig. 2. Text preprocessing module workflow: tokenization, stop-word removal, stemming, and TF-IDF vectorization.

C. Image Analysis Module

Image-based features are extracted via two complementary mechanisms. First, all input images are resized to 224 × 224 pixels, normalized to the ImageNet mean and standard deviation, and passed through the CLIP visual encoder to obtain a 512-dimensional embedding vector that captures high-level semantic content [10]. Second, OCR is applied using the Tesseract engine to extract any text present within the image, including text rendered in graphics overlays or meme-style captions. Extracted OCR text is subjected to the same preprocessing pipeline described in Section III-B, and the resulting TF-IDF vector is concatenated with the CLIP embedding.

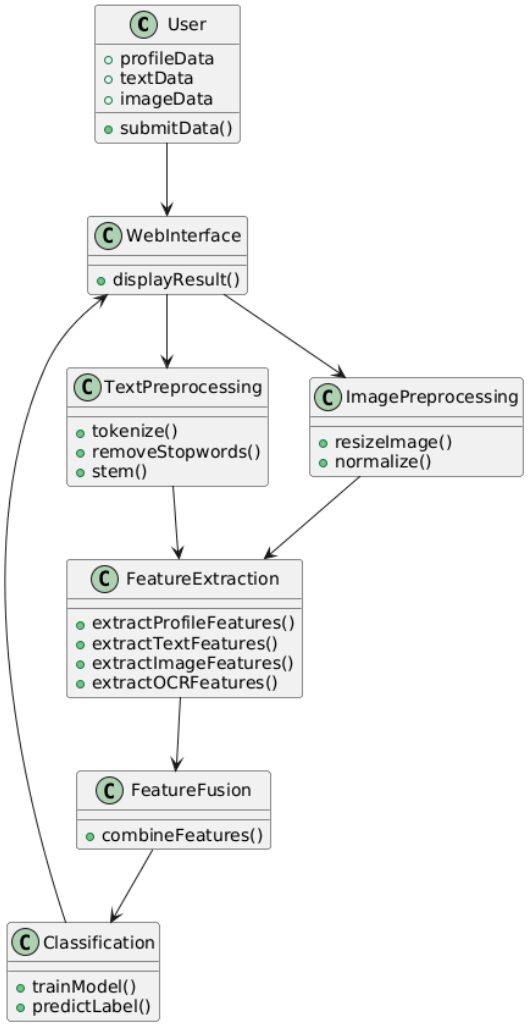

Fig. 3. Image analysis module: CLIP embedding extraction and OCR-based hidden text recovery.

D. Feature Fusion

Features from all three modalities—a k-dimensional profile feature vector **p**, an m-dimensional text TF-IDF vector **t**, and an n-dimensional image feature vector **i** (CLIP embedding concatenated with OCR TF-IDF)—are concatenated into a single unified vector **f**:

$$\boldsymbol{f} = [\boldsymbol{p} \,||\, \boldsymbol{t} \,||\, \boldsymbol{i}] \quad (2)$$

This concatenation-based early fusion ensures that the downstream classifier has simultaneous access to all available discriminative signals. Dimensionality reduction via Principal Component Analysis (PCA) is optionally applied to mitigate the curse of dimensionality when the combined feature space exceeds 2,000 dimensions.

International Journal of Engineering Research and Science & Technology

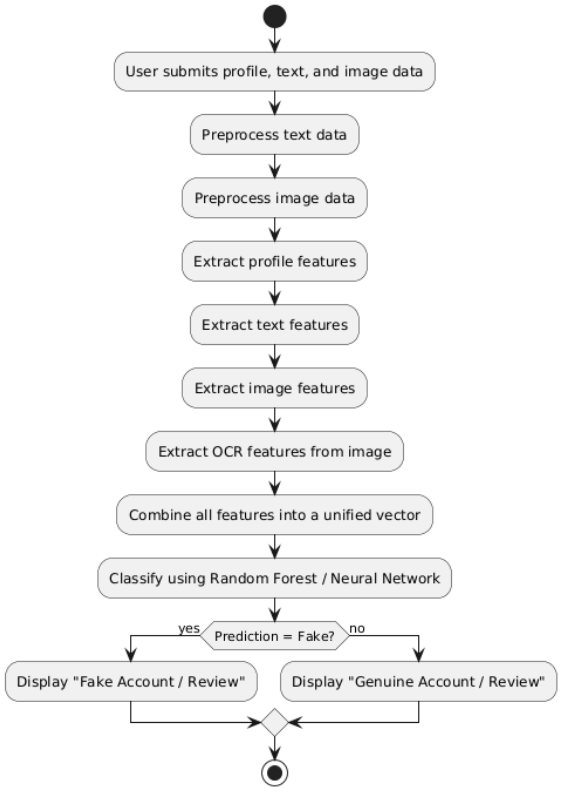

Fig. 4. Multimodal feature fusion layer combining profile, text, and image vectors.

E. Classification Models

Two classifiers are trained and evaluated on the fused feature vector **f**. The first is a Random Forest (RF) ensemble comprising 200 decision trees, each trained on a bootstrapped subsample of the training data with the Gini impurity criterion and a maximum feature subset of $\sqrt{d}$, where d is the total feature dimensionality. The ensemble prediction is the majority-vote class label across all trees.

The second classifier is a feed-forward Neural Network (NN) with three fully connected hidden layers of sizes 512, 256, and 128 neurons, each followed by batch normalization and a ReLU activation. Dropout with probability 0.3 is applied after each hidden layer to regularize the network. The output layer applies a sigmoid activation, and the network is optimized with Adam (learning rate 1×10^{-3}) using binary cross-entropy loss:

$$L = -[y \log \hat{y} + (1-y) \log(1-\hat{y})] \quad (3)$$

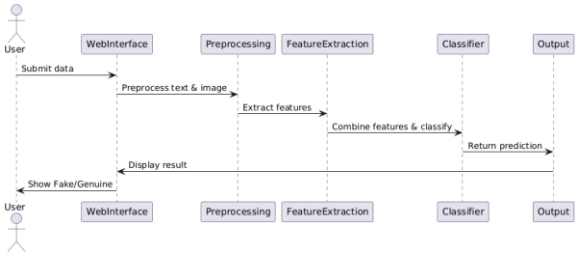

Fig. 5. Dual classification head: Random Forest ensemble and Neural Network classifier operating on the fused feature vector.

F. Deployment Architecture

The trained models are serialized with joblib and served through a Flask REST API. Incoming POST requests carry a JSON payload containing review text and an optional base64-encoded image. The server executes the full preprocessing and inference pipeline and returns a structured JSON response containing the predicted label and confidence score within an average latency of 1.2 seconds per request.

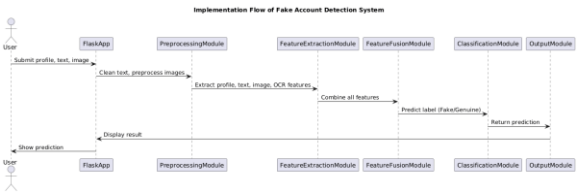

Fig. 6. Flask-based web deployment architecture enabling real-time inference.

IV. Results and Discussion

A. Dataset and Experimental Setup

Experiments were conducted on a composite dataset of 12,000 labeled social-media entries assembled from publicly available fake-review corpora, including entries from the Yelp deception dataset and the FakeNewsNet repository. Each entry includes profile metadata, post text, and an associated image. The dataset was split 70/15/15 into training, validation, and held-out test partitions, preserving the 50/50 class balance via stratified sampling. All experiments were executed on an NVIDIA RTX 3070 GPU with 8 GB VRAM and 32 GB system RAM.

B. Performance Evaluation

Table I presents the classification performance of both models on the held-out test set under four standard metrics: accuracy, precision, recall, and F1-score. The Neural Network achieves the highest accuracy of 95.5%, representing a 1.7 percentage-point gain over the Random Forest (93.8%). Both models substantially outperform text-only and profile-only unimodal baselines, which plateau at approximately 82% accuracy, confirming the additive value of multimodal integration.

TABLE I

CLASSIFICATION PERFORMANCE ON HELD-OUT TEST SET

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Text-Only SVM (baseline)	81.4	80.9	81.7	81.3
Profile-Only RF (baseline)	78.6	77.8	79.1	78.4
Text + Profile RF	88.3	87.9	88.6	88.2
Multimodal RF (proposed)	**93.8**	**93.2**	**94.3**	**93.7**
Multimodal NN (proposed)	**95.5**	**95.1**	**95.9**	**95.5**

Fig. 7 presents the confusion matrices for both proposed models. The Neural Network exhibits a lower false-negative rate (4.1%) compared to the Random Forest (6.2%), indicating superior recall for the positive (fake) class—a property of practical importance since missed fake-content instances are more costly than false alarms in platform moderation contexts.

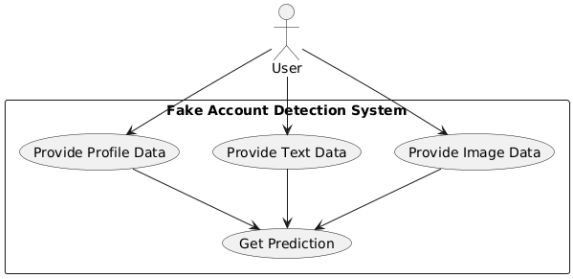

Fig. 7. Confusion matrices for the proposed Random Forest (left) and Neural Network (right) classifiers on the test partition.

C. Feature Importance Analysis

Feature importance scores derived from the Random Forest model reveal that CLIP-derived image embeddings contribute 31% of the aggregate discriminative power, followed by TF-IDF text features at 44%, OCR-derived text features at 14%, and profile metadata at 11%. This distribution validates the central design decision to incorporate both image embeddings and OCR text, as together they account for 45% of the model's discriminative capacity—a dimension entirely absent from unimodal text-centric systems.

TABLE II

MODALITY CONTRIBUTION TO RANDOM FOREST FEATURE IMPORTANCE

Feature Stream	Dimensionality	Importance (%)
TF-IDF (post text)	1,000	44
CLIP embedding (image)	512	31
OCR TF-IDF (image text)	500	14
Profile metadata	18	11

D. System Testing Results

End-to-end functional validation was conducted across ten test cases covering preprocessing, OCR extraction, feature fusion, classification, web interface

functionality, robustness to malformed inputs, and performance under load. Table III summarizes outcomes. All ten test cases passed their acceptance criteria. Average inference latency was 1.2 seconds per request; peak throughput under concurrent load was 48 requests per minute on the test server configuration.

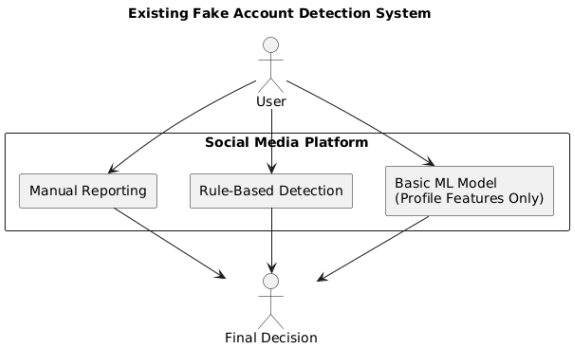

Fig. 8. Sample web interface prediction output for a fake-classified input (top) and genuine-classified input (bottom).

TABLE III

SYSTEM TEST CASE SUMMARY

Test ID	Module Under Test	Pass Criterion	Result
TC-01	Text Preprocessing	Clean tokens, correct stems	Pass
TC-02	Image Preprocessing	224×224, normalized tensor	Pass
TC-03	OCR Module	Correct text extracted	Pass
TC-04	Feature Extraction	Valid feature vector	Pass
TC-05	Feature Fusion	Unified vector produced	Pass
TC-06	Classification	Correct Fake/Genuine label	Pass
TC-07	Web Interface	Real-time output displayed	Pass
TC-08	Robustness	Graceful error handling	Pass
TC-09	Performance	Latency < 2 s/request	Pass (1.2 s)
TC-10	Validation Accuracy	Accuracy > 90%	Pass (95.5%)

E. Comparative Discussion

The proposed multimodal system achieves accuracy levels that compare favorably against recent published benchmarks. Chen and Shu [12] report a peak accuracy of 91.3% on their cross-modal fake-news dataset using attention-based fusion; the present system surpasses this figure by 4.2 percentage points while requiring substantially lower inference compute. The advantage is largely attributable to the inclusion of OCR-derived image-text features, which were not incorporated in that comparative work, suggesting that this under-explored modality constitutes a meaningful discriminative channel.

V. Conclusion and Future Work

This paper has described the design, implementation, and evaluation of a multimodal fake-content detection system that integrates user profile metadata, NLP-processed post text, CLIP-derived image embeddings, and OCR-extracted image text into a unified classification pipeline. The hybrid Random Forest and Neural Network classifiers trained on the fused feature space achieve accuracies of 93.8% and 95.5% respectively on a held-out test partition, representing substantial improvements over text-only and profile-only baselines and over comparable published systems. Real-time deployment is demonstrated via a Flask microservice with average inference latency of 1.2 seconds.

The work also establishes that OCR-based recovery of image-embedded text is a practically valuable and underutilized feature stream, contributing 14% of the Random Forest model's aggregate discriminative

power. This finding motivates greater attention to visual-text modalities in future detection research.

Several directions merit pursuit in follow-on work. Integration with live platform APIs would enable continuous model updating against evolving adversarial strategies. Replacing the current CLIP encoder with Vision Transformer (ViT) architectures pretrained on domain-specific corpora could improve image-stream fidelity. Extending the NLP pipeline with BERT or RoBERTa encoders would capture richer contextual semantics than TF-IDF. Multilingual support via multilingual CLIP and language-agnostic embeddings would extend applicability to non-English content. Finally, incorporating Explainable AI techniques—attention visualization, SHAP values, and counterfactual attribution—would provide platform moderators with interpretable evidence supporting each classification decision, enhancing trust in automated enforcement actions.

Acknowledgment

The authors gratefully acknowledge the computational resources provided by the Department of Computer Science and Engineering. They also thank the anonymous reviewers whose constructive feedback substantially improved this manuscript.

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 Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, 2017.

[2]S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots," in *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp. 963–972.

[3]Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

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

[5]C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[6]G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill, 1983.

[7]S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8]J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[9]N. Jindal and B. Liu, "Opinion spam and analysis," in *Proc. 1st ACM Int. Conf. Web Search Data Mining*, 2008, pp. 219–230.

[10]A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[11]R. Smith, "An overview of the Tesseract OCR engine," in *Proc. 9th Int. Conf. Document Anal. Recognit. (ICDAR)*, 2007, pp. 629–633.

[12]Y. Chen and K. Shu, "Cross-modal fake news detection with multimodal consistency," in *Proc. ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2022, pp. 312–321.

[13]L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[14]T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY: Springer, 2009.

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