Brand Integrity Nexus is built to help organizations protect their online identity and reduce risks linked to fake products and fraudulent digital activity

Monika Jaglan, Gobind Kumar Gautam, Sweta Sah, Vaishnavi Asthana, Shaiba Parveen Ansari
Department of Computer Engineering, MGM COET, Noida, Uttar Pradesh, India – 201309

Abstract - **In the second phase of the NISCHAL_AI project, the system was developed into a working application that monitors social media posts and identifies potentially fake or misleading brand content.**
The system retrieves social media data through Apify and analyzes captions and hashtags using a combined framework that integrates TF-IDF, Logistic Regression, keyword filtering, and hashtag-based indicators.
One of the core components is a notification module that triggers alerts when potentially harmful content is detected.
The results are displayed on a dashboard, and the system is deployed on a Hostinger VPS with the domain nishchal.online, making it accessible online. This phase demonstrates the transition from a theoretical design to a practical, real-world solution.

I. INTRODUCTION

The rapid growth of social media platforms has exposed brands to higher risks of fraudulent promotions and misleading content.

The NISCHAL_AI system automates brand monitoring by analyzing social media posts in real time. Data is collected using Apify and processed using a hybrid approach combining TF-IDF+ Logistic Regression, keyword filtering, and hashtag-based signals.

A dedicated image analysis pipeline is implemented, where YOLOv8 performs object detection and EfficientNet handles image classification. The model assigns a probability score to each post, which is then used to determine whether the content is genuine or deceptive.

An important feature of the system is the alert mechanism, which sends notifications via Gmail and Twilio SMS when suspicious content is detected, enabling quicker response and monitoring.

The results are displayed through an interactive dashboard, making the system practical, efficient, and deployable.

Objectives:

1. To develop a real-time system for monitoring brand-related social media content
2. To detect counterfeit or misleading posts using a hybrid approach (TF-IDF + Logistic Regression + keyword filtering)
3. To incorporate hashtag-based signals for improving detection accuracy
4. To integrate image-based verification using YOLOv8 and EfficientNet
5. To generate a confidence score for classifying posts as real, fake, or uncertain
6. To implement an alert system using Gmail and Twilio SMS for notifying suspicious activity
7. To design an interactive dashboard for visualization and monitoring
8. To deploy the system on a VPS server (nishchal.online) for real-world accessibility

Research Questions:

1. How effective is the hybrid model in detecting fake or misleading social media posts?
2. Does combining keyword filtering with machine learning improve prediction accuracy?
3. What role do hashtags play in identifying suspicious or promotional content?
4. How accurate are YOLOv8 and EfficientNet in detecting counterfeit-related image features?
5. How does combining text and image analysis improve overall detection

reliability?

6. What is the impact of confidence thresholds on classification performance?
7. How effective is the alert system (Gmail + Twilio) in enabling real-time monitoring?
8. Can the system perform efficiently when deployed in a real-world environment?

II. LITERATURE REVIEW

Previous research indicates that identifying misleading or counterfeit content on social platforms requires integrating both textual and visual analysis techniques. Traditional NLP models like BERT and DistilBERT are effective in understanding captions and detecting hidden intent, while computer vision models such as ResNet and Vision Transformers help identify visual inconsistencies in product images.

Multimodal models like CLIP have been widely used to compare image–text similarity, making them useful for detecting mismatched or misleading posts. Findings suggest that multi-feature approaches outperform single-method techniques in terms of detection accuracy.

However, most existing approaches focus on offline analysis and require high computational resources, making real-time deployment difficult. In this project, a more practical approach is adopted by using lightweight models and hybrid techniques to achieve efficient and deployable counterfeit detection.

III. METHODOLOGY

3.1. Data Collection

Social media data is collected using Apify, which provides structured information such as captions, hashtags, and engagement details.

3.2. Data Preprocessing

The collected data is cleaned and standardized before analysis.

- Text data is normalised by converting it to lowercase and removing noise such as URLs, emojis, and irrelevant symbols. Relevant hashtags are extracted and standardized.
- While the cleaned text is transformed into TF-IDF features for further analysis.

This step ensures consistency and improves model performance.

3.3. Text Analysis and Scoring

A hybrid approach is used for text-based detection:

- Textual patterns are evaluated using a TF-IDF representation combined with a Logistic
  Regression model, contributing significantly to the overall scoring.
- A predefined keyword filtering mechanism identifies suspicious terms (e.g., “replica” or “first copy”) commonly associated with counterfeit content.
- Additional signals are derived from hashtags

These components are combined to generate a fake probability score for each post.

3.4. Image-Based Detection

For image verification, a separate detection pipeline is used:

- YOLOv8 is applied to identify objects and extract relevant visual patterns from images.
- EfficientNet is utilized to classify images based on learned visual features.

This module is mainly used for user-uploaded images, reducing system load during continuous monitoring.

3.5. Classification and Decision Making

Based on the combined score:

- Posts are classified as Real, Fake, or Uncertain
- Confidence thresholds are applied to improve accuracy

3.6. Alert System

When a post is identified as suspicious notifications are delivered through email (Gmail) and SMS services (Twilio).
This enables real-time monitoring and quick response.

3.7. Visualization and Dashboard

The processed results are displayed on a dashboard showing:

- Total mentions
- Fake vs real posts
- Brand reputation metrics
- Live feed of analyzed posts

3.8. Deployment

The system is deployed on a Hostinger VPS (KVM1 plan) and connected to the domain Nishchal.online, making it accessible online.

Expected Outcome

1. Successful implementation of a real-time system for monitoring social media posts
2. Accurate classification of posts into real, fake, or uncertain categories
3. Generation of a confidence score based on hybrid text and image analysis
4. Efficient detection using TF-IDF + Logistic Regression, keyword filtering, YOLOv8, and EfficientNet
5. Real-time alert system using Gmail and Twilio SMS for suspicious content
6. Interactive dashboard displaying brand insights and live feed of posts
7. Stable deployment of the system on a VPS (nishchal.online)
8. Improved performance through optimized and lightweight model usage

Architecture Overview

1. Data Ingestion Layer
   - Social media data is collected using Apify
   - Extracts captions, hashtags, and metadata
2. Preprocessing Layer
   - Cleans and normalizes text data
   - Extracts hashtags and prepares input for analysis
3. Text Analysis Layer
   - Classification is carried out using TF-IDF features combined with a Logistic Regression model.
   - A filtering mechanism for detecting suspicious keywords.
   - Incorporates hashtag-based signals
4. Image Detection Layer
   - Uses YOLOv8 for object detection
   - Uses EfficientNet for image classification
   - Applied mainly to user-uploaded images
5. Scoring & Decision Layer
   - Combines all signals to generate a confidence score
   - Classifies posts as Real, Fake, or Uncertain using thresholds
6. Alert System Layer
   - The system triggers alerts via Gmail.
   - Twilio to ensure timely user notification.
7. Visualization Layer
   - Dashboard displays:
     - Live feed
     - Fake vs real posts
     - Brand reputation metrics
8. Deployment Layer
   - Hosted on Hostinger VPS (KVM1)
   - Accessible via nishchal.online

Implementation Tool

- Apify is used for collecting social media data such as captions and hashtags
- FastAPI is used to build the backend and handle API requests
- HTML, CSS, and JavaScript are used to create the dashboard interface
- TF-IDF and Logistic Regression (scikit-learn) are used for text classification
- DistilBERT is used for additional text understanding
- YOLOv8 is used for object detection in images
- EfficientNet is used for image classification
- SQLite is used to store user data, posts, and results
- Communication with users is handled through email and SMS integration.
- The entire system is developed primarily using Python.
- Libraries such as NumPy, Pandas, and OpenCV are used for data processing
- Uvicorn is used to run the FastAPI

server
- The system is deployed on Hostinger VPS (KVM1)
- The application is accessible through the domain nishchal.online

Counterfeit Detection — Summary

The proposed system integrates textual analysis, keyword filtering, and image-based verification to detect misleading content on social media platforms.Text-based detection using TF-IDF + Logistic Regression achieves an accuracy of approximately 85–88%, effectively identifying suspicious captions

- Keyword-based filtering improves detection by capturing common counterfeit terms, contributing to overall system reliability
- Image-based detection using YOLOv8 and EfficientNet provides accuracy in the range of 88–92%, helping identify visual inconsistencies
- Combined hybrid approach improves overall detection performance to approximately 90–92% accuracy
- The model achieves a precision of approximately 90%, which helps minimize false positive detections.
- Recall values between 91% and 93% indicate that the system successfully captures
  the majority of fraudulent posts.
- F1 Score is around 91–92%, indicating a balanced performance between precision and recall

The system uses confidence thresholds to classify posts:

- Below threshold → Real
- Above threshold → Fake
- Mid-range → Uncertain

Overall, results show that combining text, keyword, and image analysis provides more reliable detection compared to using a single method, while maintaining efficiency for real-time deployment.

IV. BRAND AUTHENTICITY-SUMMARY

Brand authenticity significantly influences user perception and trust in digital environments. Genuine and consistent content increases user confidence, while misleading or counterfeit posts can negatively impact brand reputation.

In this system, brand authenticity is evaluated by analyzing social media posts using a combination of text and image-based techniques. Posts classified as fake or suspicious reduce the overall brand reputation score, while genuine posts contribute positively.

The dashboard reflects this through real-time metrics such as fake vs real post distribution and overall reputation trends. Additionally, the alert system helps in identifying harmful content early, allowing quicker action to protect brand integrity.

Overall, maintaining a higher proportion of authentic content leads to improved trust, engagement, and brand value, making automated monitoring essential for modern digital platforms.

Note on Model Selection and Blockchain Usage

In the initial phase of the project, advanced models such as BERT, CLIP, Vision Transformers (ViT), and blockchain-based storage were considered as part of the system design.

However, during implementation, these components were not fully used due to practical constraints. The primary challenges included high memory consumption, increased processing time, and deployment limitations on the available VPS environment. These factors made it difficult to achieve stable real-time performance using heavy models.

Similarly, while blockchain technology was initially proposed for secure and tamper-proof storage of suspicious posts, it was not implemented in this phase due to its added complexity and integration overhead.

To ensure better performance and successful deployment, the system was optimized using lightweight and efficient alternatives, including DistilBERT, TF-IDF with Logistic Regression, YOLOv8, and EfficientNet.

This approach allowed the system to remain practical, responsive, and deployable while still maintaining effective detection capability.

V. CONCLUSION

In Phase 2, the NISCHAL_AI proposed system has evolved from a conceptual framework into a fully functional and deployable solution. The system effectively identifies counterfeit and misleading content across social media platforms using a hybrid analytical approach.

Key features such as real-time monitoring, dashboard visualization, and alert mechanisms (email and SMS) make the system practical for real-world usage. Deployment on a VPS with a custom domain further demonstrates its scalability and usability.

Overall, the project highlights the importance of optimizing models and designing efficient pipelines to achieve reliable performance in real-time environments.

REFERENCE

[1] Devlin, J. et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

[2] Sanh, V. et al. (2019). *DistilBERT: A distilled version of BERT*

[3] He, K. et al. (2016). *Deep Residual Learning for Image Recognition (ResNet)*

[4] Radford, A. et al. (2021). *CLIP: Learning Transferable Visual Models from Natural Language Supervision*

[5] Redmon, J. et al. (2016). *YOLO: You Only Look Once – Unified, Real-Time Object Detection*

[6] Tan, M. & Le, Q. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*

[7] Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python*

[8] Apify. (2024). *Web Scraping and Automation Platform*

[9] Twilio. (2024). *Cloud Communications Platform for SMS Services*

[10] FastAPI Documentation. (2024). *Modern Web Framework for APIs*