International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 08 August-2026 | Impact Factor: 3.5

A Modular Multimodal Framework for Automated Misinformation Investigation: Integrating OCR, NLP, Vision-Language Captioning, and Speech Analysis for Evidence-Based Trust Scoring

Author: **Shivani Tiwari**
Military College of Telecommunication Engineering, Mhow, India
Shivanitiwari.1122@gmail.com

How to Cite this Article:

Tiwari, S. (2026). A Modular Multimodal Framework for Automated Misinformation Investigation: Integrating OCR, NLP, Vision-Language Captioning, and Speech Analysis for Evidence-Based Trust Scoring. International Journal of Creative and Open Research in Engineering and Management, 02(8), 1-9. https://doi.org/10.55041/ijcope.v2i8.147

https://doi.org/10.55041/ijcope.v2i8.147

***Abstract*—** The rapid spread of manipulated images, doctored videos, and misleading text across digital platforms has created a pressing need for automated tools that can assist human fact-checkers rather than replace them. This paper presents the design and implementation of a modular, multimodal misinformation investigation platform that combines optical character recognition (OCR), natural language processing (NLP), vision-language image captioning, and automatic speech recognition (ASR) within a single evidence-fusion pipeline. The system extracts on-screen text, spoken audio, visual content descriptions, and linguistic sentiment/clickbait indicators from a submitted image or video, and combines these independent signals into a composite trust score and human-readable explanation through a rule-based fusion engine. The platform is implemented as a FastAPI backend with a MySQL relational store, JSON Web Token (JWT) based authentication, and a React single-page-application frontend that supports investigation history, analytics, and report export. We describe the architecture, the responsibilities of each processing module, the design of the evidence-fusion scoring function, and the security and data-persistence layers. We further discuss the current limitations of the rule-based fusion approach, including its sensitivity to keyword-based clickbait detection and its reliance on unimodal confidence rather than learned cross-modal weighting, and we outline directions for extending the system toward a trained fusion model and stronger authentication practices.

***Keywords*—** *misinformation detection, multimodal fusion, optical character recognition, natural language processing, vision-language models, speech recognition, trust scoring, FastAPI, digital forensics.*

I. Introduction

The proliferation of short-form video, screenshots, and voice content on social platforms has outpaced the ability of manual fact-checking teams to verify claims before they spread. A single piece of misleading media often carries several independent signals of manipulation or manipulation intent: overlaid text designed to provoke a reaction, a spoken narration whose claims are inconsistent with the visible scene, emotionally charged or clickbait phrasing, and a visual composition that a caption model would describe differently than the accompanying text suggests. Treating these signals in isolation discards evidence that becomes meaningful only when combined.

This paper describes a multimodal investigation platform built to give an investigator a single, explainable trust assessment that is assembled from several independent AI components: an OCR engine for on-screen text, an NLP engine for sentiment and clickbait phrasing, a vision-language captioning model for describing visual content, and a speech-to-text engine for spoken audio in video evidence. Rather than producing a single opaque classification, the platform is designed around an evidence-fusion approach: each module contributes an observation, and a fusion engine combines these observations into a bounded trust score together with a plain-language explanation of which factors reduced or supported the score. This design choice reflects the intended use case — assisting a human investigator rather than issuing an unreviewable automated verdict.

The remainder of this paper is organized as follows. Section II reviews related approaches to automated misinformation detection. Section III describes the overall system architecture. Section IV details the methodology of each processing module and the fusion scoring function. Section V discusses implementation aspects, including authentication and persistence. Section VI walks through the system's behavior on a representative investigation case. Section VII discusses limitations, and Section VIII concludes with directions for future work.

II. Related Work

Automated misinformation detection has generally been approached from three angles: text-based claim analysis, image/video forensics, and, more recently, multimodal fusion. Text-based approaches typically rely on linguistic cues such as sentiment, subjectivity, and stylistic markers associated with clickbait or sensationalized writing [1], [2]. Image and video forensics approaches focus on detecting pixel-level manipulation artifacts, splicing, or generative-model fingerprints, but these techniques generally do not reason about whether the accompanying text or narration is consistent with the depicted content. Vision-language models such as BLIP [3] were originally developed for image captioning and visual question answering, but their descriptive captions can also serve as an independent, human-readable check on whether a claim made in accompanying text matches what is actually visible in an image. Automatic speech recognition systems such as Whisper [4] make it practical to bring spoken narration in video evidence into the same textual analysis pipeline used for on-screen text and captions, an integration step that many single-modality misinformation tools omit. General-purpose NLP pipelines built on spaCy [5] and transformer-based sentiment classifiers such as DistilBERT [6] provide lightweight, well-understood building blocks for the linguistic layer of such a system. Broad surveys of the fake-news detection literature [7], [8] note that multimodal and fusion-based approaches remain comparatively underexplored relative to single-modality text or image classifiers, which motivates the fusion-oriented design adopted in this work.

III. System Architecture

The platform follows a layered client-server architecture. The frontend is a React single-page application (built with react-router-dom) exposing distinct views for authentication (Login/Register), image investigation, video investigation, investigation history, analytics, report generation, and per-investigation detail pages, all wrapped in a shared MainLayout component. The backend is implemented in Python using FastAPI, and is organized as a set of routers mounted on a single application instance: authentication, OCR, image analysis, video analysis, analytics, history, reports, export, investigation management, search, and notifications. Cross-Origin Resource Sharing (CORS) is configured explicitly to permit the local development frontend origins, and uploaded evidence files are served through a dedicated static-files mount.

Persistence is handled through SQLAlchemy's ORM layer against a MySQL database, using a session-per-request pattern (a generator-based get_db dependency that yields a session and guarantees it is closed after the request completes). An Investigation model stores, per submitted file, the filename, storage path, inferred media type (image or video, determined from the file extension), the fusion engine's prediction and trust score, the extracted OCR text, a string representation of the vision result, the detected sentiment label, the clickbait flag, and the human-readable explanation returned by the fusion engine.

Figure 1 (conceptual, not reproduced here) summarizes the data flow: a submitted image or video first passes through media-specific preprocessing (frame extraction for video), then through the OCR, NLP, vision, and — for video — speech modules in parallel or sequence, after which the fusion engine combines their outputs into a single scored, explained result that is persisted and returned to the frontend for display and later retrieval.

IV. Methodology

A. Text Extraction (OCR)

On-screen text is extracted using EasyOCR [9], a deep-learning-based OCR engine, configured for English-language recognition on CPU. Each detected text region's recognized string is concatenated into a single extracted-text string that

is passed downstream to the NLP module and used as one of the fusion engine's evidence signals. The presence or absence of any readable text is itself treated as a weak evidence signal: media with no extractable text is scored as offering less textual evidence to corroborate or contradict.

B. Linguistic Analysis (NLP)

The NLP engine performs three linguistic analyses on the combined text (OCR output concatenated with any transcript, for video). First, spaCy's small English pipeline [5] performs named-entity recognition, extracting entities and their labels so that an investigator can see which people, organizations, locations, or other entities are referenced. Second, a DistilBERT-based sentiment classifier fine-tuned on the Stanford Sentiment Treebank (SST-2) [6] assigns a POSITIVE/NEGATIVE sentiment label with a confidence score. Third, a lexicon-based clickbait detector flags text containing any of a curated list of sensationalized terms (e.g., "breaking", "shocking", "urgent", "exclusive", "viral", "secret", "truth", "warning"). Both the sentiment label and the clickbait flag feed directly into the fusion engine's scoring function.

C. Visual Content Analysis

Visual content is described using BLIP [3], a vision-language model that generates natural-language image captions. For video evidence, the first decodable frame is extracted with OpenCV and written to a temporary file before captioning; the caption generation uses beam search (five beams) with a repetition penalty and both minimum and maximum new-token constraints to produce a caption of bounded length and diversity. The resulting caption serves two purposes: it gives the investigator an independent, human-readable description of the visual content that can be compared against any textual claim, and its non-emptiness is treated by the fusion engine as evidence that visual analysis succeeded.

D. Speech-to-Text (Video Only)

For video submissions, spoken audio is transcribed using OpenAI's Whisper model [4] (base configuration), loaded once at process start. The resulting transcript is concatenated with the OCR text before being passed to the NLP module, so that sentiment and clickbait analysis reflect both on-screen text and spoken narration. Transcription failures are handled defensively: an exception during transcription yields an empty transcript rather than aborting the investigation pipeline.

E. Evidence Fusion and Trust Scoring

The core novelty of the platform lies in its fusion engine, which combines the OCR, NLP, and vision outputs into a single bounded trust score using an explainable, rule-based deduction model rather than a black-box classifier. Scoring begins at a maximum of 100 points. Ten points are deducted when no readable text is detected, reflecting reduced textual evidence; otherwise, the presence of extracted text is logged as a positive explanatory note. Ten points are deducted when the NLP module reports negative sentiment, and a further twenty points are deducted when clickbait language is detected, reflecting the stronger association between clickbait phrasing and low-credibility content. Fifteen points are deducted when the vision module returns no usable caption, and the presence of a caption is otherwise logged as a positive note. The resulting score is clipped to the [0, 100] range and mapped to one of three categorical predictions: "Likely Genuine" (score $\geq$ 70), "Suspicious" (40 $\leq$ score $<$ 70), or "Likely Misinformation" (score $<$ 40). Alongside the numeric score, the fusion engine returns an ordered list of plain-language explanation strings, so that every deduction is traceable to a specific piece of evidence — a deliberate design choice to keep the system auditable by a human investigator rather than presenting an unexplained number.

F. Authentication and Security

User sessions are authenticated using JSON Web Tokens signed with the HS256 algorithm via the python-jose library, with a configurable expiry (currently 60 minutes). Passwords are intended to be hashed using passlib with bcrypt, consistent with the project's declared dependencies. The current JWT signing secret is defined as a static string constant in source code rather than being loaded from an environment variable or secrets manager; Section VII discusses this as a limitation requiring remediation before any production deployment.

G. Persistence and Investigation Records

Once a fusion result is produced, the investigation service classifies the submitted file's media type from its extension, constructs an Investigation ORM record capturing the filename, storage path, media type, prediction, trust score, OCR text, a string representation of the vision result, sentiment label, clickbait flag, and explanation text, and commits it to the database within a try/except/finally block that rolls back on failure and always closes the session. This record subsequently backs the history, analytics, reporting, and export features exposed through the frontend.

V. Representative Investigation Walkthrough

To illustrate the pipeline's behavior, consider a short video submitted for investigation. The video engine first attempts to decode the initial frame with OpenCV; if decoding fails, the pipeline short-circuits and returns an "Error" prediction with a zero trust score and an explanation noting that the video could not be decoded, rather than attempting to fabricate a result from partial data. If the frame decodes successfully, it is written to a temporary JPEG file, OCR is run on that frame, Whisper transcribes the full audio track, the OCR text and transcript are concatenated, spaCy and the sentiment/clickbait detectors analyze the combined text, BLIP captions the frame, and the fusion engine combines all three signals into a final score and explanation. The temporary frame file is removed in a finally block regardless of whether the intermediate steps succeeded, avoiding accumulation of orphaned temporary files across repeated investigations. The final result — including the raw OCR text, transcript, caption, sentiment, clickbait flag, prediction, score, and explanation — is both returned to the frontend for

immediate display and persisted for later retrieval through the history and analytics views.

VI. Discussion

The rule-based fusion design has two notable strengths for the investigative use case this platform targets. First, every score is fully explainable: an investigator can see exactly which of three evidence categories (textual presence, sentiment/clickbait, visual analysis availability) contributed to a deduction, which is important when a human is expected to review and potentially override an automated assessment. Second, the modular separation between OCR, NLP, vision, and speech components means any individual module can be upgraded — for example, replacing the lexicon-based clickbait detector with a learned classifier, or replacing BLIP with a larger vision-language model — without requiring changes to the fusion logic or the rest of the pipeline.

At the same time, the current fusion function has clear limitations that should be weighed against these strengths, which are discussed in detail in Section VII.

VII. Limitations

Several limitations of the current design should be addressed before broader deployment. First, the fusion engine's weights (10, 10, 20, and 15 points for the four respective deductions) are fixed and hand-chosen rather than learned from labeled data; this makes the scoring function easy to explain but not empirically calibrated against ground-truth misinformation labels. Second, the clickbait detector relies on a fixed English-language keyword list, which is unlikely to generalize across languages, evolving slang, or adversarially reworded clickbait phrasing. Third, the fusion engine does not currently perform cross-modal consistency checking — for example, comparing whether the BLIP caption of a frame actually contradicts a claim made in the OCR text or transcript — even though the modality outputs required for such a check are already computed by the pipeline. Fourth, the JWT signing key is presently a hard-coded constant rather than an externally managed secret, which is a security risk that should be corrected by loading the key from environment configuration before any deployment beyond local development. Finally, video analysis is currently based on a single decoded frame rather than sampling across the video timeline, which may miss manipulated or misleading content that appears only in later frames.

VIII. Conclusion and Future Work

This paper presented the design of a modular, multimodal misinformation investigation platform that fuses OCR, NLP, vision-language captioning, and speech-to-text signals into a single explainable trust score, implemented on a FastAPI/MySQL backend with a React frontend and JWT-based authentication. The system's explicit, rule-based fusion approach favors auditability over raw predictive accuracy, which is well suited to an investigator-assistance tool. Future work includes: (1) replacing the fixed-weight fusion function with a model trained on labeled misinformation datasets while retaining a human-readable explanation layer; (2) adding explicit cross-modal contradiction detection between visual captions and textual/spoken claims; (3) extending video analysis to multi-frame or scene-level sampling rather than a single initial frame; (4) migrating the JWT signing secret and other credentials to environment-based secret management; and (5) conducting a formal evaluation against a labeled misinformation benchmark to quantify precision, recall, and calibration of the trust score, which the current implementation does not yet include.

References

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146–1151, 2018.

[2] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 2931–2937.

[3] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in Proc. Int. Conf. Machine Learning (ICML), 2022.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," arXiv preprint arXiv:2212.04356, 2022.

[5] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017. [Online software].

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[7] 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.

[8] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," ACM Computing Surveys, vol. 53, no. 5, pp. 1–40, 2020.

[9] JaidedAI, "EasyOCR: Ready-to-use OCR with 80+ supported languages," GitHub repository, 2020. [Online software].

[10] T. Ramalho et al., "FastAPI: A modern, fast web framework for building APIs with Python," 2018–present. [Online software].