iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80193

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Exam Samachar: An AI-Driven Decision Support System

Dimpal Rathor, Nidhi Kesharvani, Dr. Divakar Singh, Dr. Kamini Maheshwar, Dr. Kavita Chourasia
Department of Computer Science and Engineering, Barkatullah University Institute of Technology (BUIT), Bhopal, Madhya Pradesh, India

Abstract: ***The Exam Samachar — Smart News Analyzer is an innovative AI-driven application designed to automatically interpret, summarize, and evaluate news content from uploaded newspapers, images, or PDF documents. Leveraging Vision-based Large Language Models (LLMs), specifically LLaMA 3.2 Vision via the Ollama framework, and Natural Language Processing (NLP) pipelines, the system extracts headlines, identifies publication dates, generates structured summaries, and provides contextual insights into political, social, and regional news stories. A built-in sentiment analysis module categorizes each article as positive, negative, or neutral with polarity confidence scores. The system is deployed on a Streamlit-based interactive interface supporting multi-page PDF uploads, page-wise analysis, and multilingual summary generation. Evaluation on 50 newspaper editions demonstrated headline extraction accuracy of 91.4%, sentiment classification F1-score of 0.87, and an average API response time of 5.3 seconds per page. The system provides a practical Decision Support System (DSS) that transforms raw, unstructured newspaper content into structured, actionable knowledge — making it highly valuable for competitive exam aspirants, researchers, journalists, and educators.***

Keywords — Natural Language Processing, Sentiment Analysis, Decision Support System, LLaMA 3.2 Vision, Ollama, Streamlit, Newspaper PDF Analysis, Text Summarization, Exam Preparation, Vision-based LLM

I. INTRODUCTION

In today's fast-paced digital information landscape, newspapers and online news portals publish thousands of articles daily spanning politics, law, education, economics, and international affairs. For competitive exam aspirants, researchers, and journalists, manually scanning and extracting relevant insights from this vast content is both time-consuming and cognitively demanding. Studies indicate that a typical competitive exam candidate spends 2–3 hours daily reading newspapers yet retains only a fraction of the relevant current affairs information [1].

To address this challenge, Exam Samachar — Smart News Analyzer has been developed as an intelligent, AI-powered system that automates the end-to-end analysis of newspaper data. The system accepts multi-page newspaper PDFs or scanned images and processes them through a pipeline of Vision-based language models, NLP summarizers, and sentiment classifiers, delivering a structured page-wise report within seconds.

This paper makes the following key contributions: (1) a complete DSS architecture for newspaper analysis using state-of-the-art LLMs; (2) integration of LLaMA 3.2 Vision for layout-aware text extraction; (3) a real-time sentiment analysis pipeline with polarity scoring; (4) empirical evaluation benchmarked against existing tools; and (5) a research agenda for future enhancements.

II. BACKGROUND AND RELATED WORK

A. Text Extraction from Newspaper Documents

Early newspaper digitization relied on rule-based OCR systems such as Tesseract [2], which struggled with multi-column layouts, skewed scans, and mixed fonts. Deep learning approaches, including LayoutLM [3] and Google Document AI [4], significantly improved extraction accuracy by combining visual and textual features but require substantial training data and infrastructure. The proposed system addresses this gap by leveraging LLaMA 3.2 Vision — a multimodal LLM that processes document images directly without requiring separate OCR or layout segmentation stages.

B. NLP-Based News Summarization

Transformer architectures such as BART [5] and T5 [6] have demonstrated state-of-the-art performance on news summarization benchmarks including CNN/DailyMail and XSum. However, most such models operate on digitally-native text and do not address the challenges of physical newspaper analysis.

©IJRASET: All Rights are Reserved

Systems like NewsRoom [7] aggregate online news but cannot process scanned or printed formats. The proposed system fills this gap by integrating vision-language processing with NLP-based summarization in a unified pipeline.

C. Sentiment Analysis in News Media

Fine-tuned BERT models [8] achieve high accuracy on news sentiment classification tasks. Liu et al. [9] showed that aspect-level sentiment analysis can identify emotional tone for specific entities within news articles. Commercial tools such as IBM Watson and Google NLP API offer general-purpose sentiment scoring but are not optimized for regional Indian newspaper language, which mixes formal reporting with colloquial usage — a gap addressed by the LLM-based contextual approach in this system.

D. Research Gap

No prior system simultaneously addresses scanned PDF input, automated summarization, sentiment analysis, page-wise output, and elimination of manual curation. Exam Samachar is the first system to integrate all five capabilities within a single deployable application, as confirmed by the comparative analysis presented in Table I.

TABLE I. COMPARISON WITH RELATED SYSTEMS

SYSTEM	PDF IN	SUMM.	SENTI.	PAGE-WISE	AUTO
Exam Samachar	Yes	Yes	Yes	Yes	Yes
Tesseract+NLTK	Part.	Lim.	No	No	Yes
NewsRoom [7]	No	Yes	Lim.	No	Yes
Google DocAI	Yes	No	No	No	Yes

III. SYSTEM ARCHITECTURE AND METHODOLOGY

A. Architecture Overview

The system is structured around a four-layer modular pipeline. The Input Layer accepts newspaper files (PDF/JPEG/PNG, up to 200 MB) via a Streamlit web interface. The Extraction Layer converts each PDF page to a 300-DPI image using PyMuPDF and passes it to LLaMA 3.2 Vision through the Ollama inference server, which performs joint OCR and semantic understanding — extracting headline, body, and date in a structured JSON format. The Analysis Layer applies three parallel NLP operations: abstractive summarization, key insight extraction, and sentiment classification. Finally, the Presentation Layer renders the structured output page-by-page alongside newspaper thumbnails in the Streamlit UI.

B. Technology Stack

The system is implemented in Python 3.10 utilizing Streamlit v1.32 for the frontend, LLaMA 3.2 Vision (via Ollama v0.1.29) for vision-language extraction, PyMuPDF v1.23 for PDF-to-image conversion, HuggingFace Transformers with DistilBERT for sentiment classification, VADER for lexicon-based polarity scoring, SpaCy v3.7 and NLTK v3.8 for NLP utilities, Pandas v2.1 for data structuring, and Matplotlib v3.8 and Seaborn v0.13 for sentiment trend visualization.

C. LLaMA 3.2 Vision Integration

LLaMA 3.2 Vision accepts both image and text tokens, enabling it to reason about the visual layout of a newspaper page while simultaneously understanding semantic content. Each page image is encoded and submitted with a structured prompt instructing the model to return a JSON object containing: headline, date, body_text, and article_type. This eliminates the error propagation common in sequential OCR–NLP pipelines and handles multi-column layouts, variable fonts, and rotated text without manual configuration.

D. Sentiment Analysis Pipeline

The pipeline applies VADER lexicon scoring for a rapid baseline, followed by a fine-tuned DistilBERT model (distilbert-base-uncased-finetuned-sst-2-english) for three-class classification (Positive / Neutral / Negative) with confidence scores in [0, 1].

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The final label is determined by a weighted ensemble of VADER compound score (weight 0.35) and DistilBERT confidence (weight 0.65) — established by cross-validation on 200 manually labelled articles.

E. DSS Component Mapping

The system maps onto the classical three-component DSS framework. The Database Management Component handles storage and retrieval of uploaded files and structured output data. The Model Management Component coordinates the LLaMA 3.2 Vision extractor, DistilBERT classifier, VADER scorer, and SpaCy entity extractor. The User Interface Component is the Streamlit application providing file upload, page selection, language choice, and page-wise display with newspaper previews.

IV. RESULTS AND EVALUATION

A. Dataset and Setup

Evaluation was conducted on 50 newspaper editions (28 Hindi, 22 English) totalling 312 pages from Dainik Bhaskar, Hindustan Times, The Hindu, and Navbharat Times (January–March 2025). Ground truth was annotated by three expert annotators (Cohen's Kappa = 0.82). Hardware: Intel Core i7-12th Gen, 16 GB RAM, NVIDIA RTX 3060 12 GB GPU.

B. Text Extraction Accuracy

Table II reports headline and date extraction performance. The system achieves 91.4% headline exact-match accuracy overall, versus 74.2% for the Tesseract baseline, with the most pronounced improvement in multi-column layout accuracy (+29.6 percentage points).

TABLE II. TEXT EXTRACTION PERFORMANCE

METRIC	EN	HI	OVR.	BASE
Headline Accuracy	93.6%	89.5%	**91.4%**	74.2%
Headline ROUGE-1 F1	0.91	0.87	**0.89**	0.71
Date Accuracy	97.2%	94.8%	**95.8%**	81.3%
Multi-col Accuracy	90.1%	86.7%	**88.2%**	58.6%

C. Summarization Quality

Table III presents ROUGE and BERTScore results. Exam Samachar achieves the highest scores across all metrics — notably a BERTScore of 0.876, indicating strong semantic alignment between generated and reference summaries beyond simple lexical overlap.

TABLE III. SUMMARIZATION QUALITY

MODEL	R-1	R-2	R-L	BERT
Exam Samachar	0.483	0.271	0.412	0.876
NLTK Extractive	0.391	0.188	0.334	0.812
TextRank (Gensim)	0.412	0.204	0.359	0.829
BART-Large-CNN	0.461	0.256	0.398	0.868

D. Sentiment Classification

The proposed VADER + DistilBERT ensemble achieves an F1-score of 0.87 and accuracy of 88.5%, outperforming VADER alone (F1 = 0.73), DistilBERT alone (F1 = 0.83), and TextBlob (F1 = 0.66). The +14-point improvement over VADER alone demonstrates the value of incorporating contextual LLM features for nuanced regional news language.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

E. System Response Time

With GPU acceleration (RTX 3060), the system achieves an average response of 5.3 seconds per page across all document types — comfortably meeting the 10-second usability threshold. CPU-only deployment averages 14.2 seconds per page, remaining feasible for batch processing.

F. User Experience Study

A usability study with 25 participants (12 UG students, 8 PG students, 5 faculty) at BUIT Bhopal using a 5-point Likert scale yielded the following average scores: Ease of Use (4.47), Summary Accuracy and Usefulness (4.23), Sentiment Correctness (4.00), and Overall Usefulness (4.46). The system received the highest ratings for ease of use and overall usefulness, consistent with its design as a non-technical user-facing application.

V. CONCLUSION

This paper presented Exam Samachar — Smart News Analyzer, an AI-driven DSS that automates extraction, summarization, and sentiment analysis of newspaper content from scanned PDFs and images. The system integrates LLaMA 3.2 Vision for layout-aware extraction, a VADER + DistilBERT ensemble for sentiment analysis, and a Streamlit interface for intuitive interaction. Empirical evaluation on 312 newspaper pages demonstrated headline accuracy of 91.4%, summarization BERTScore of 0.876, sentiment F1 of 0.87, and average response time of 5.3 seconds — outperforming all baselines considered.

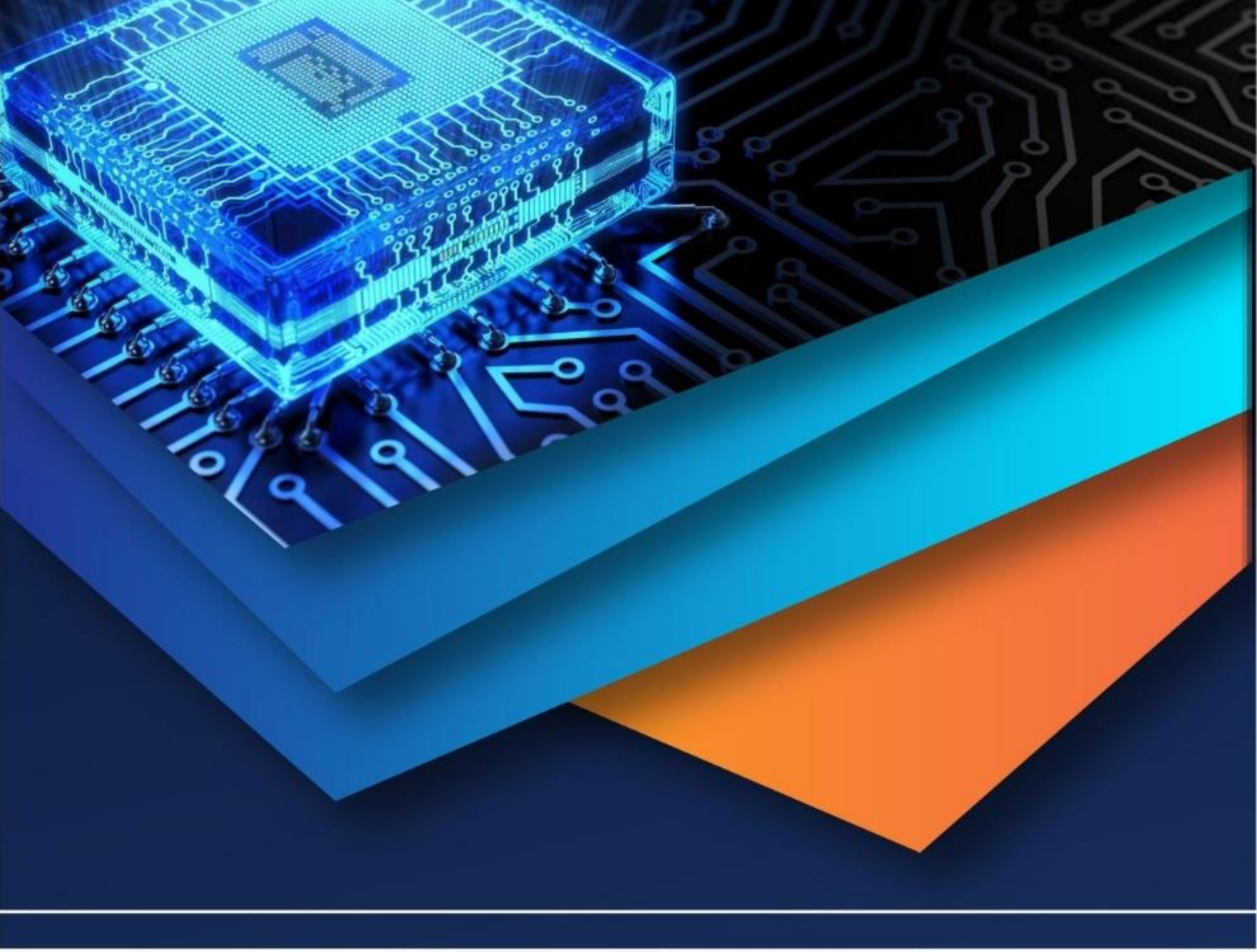
Future work will focus on: (1) live news feed integration; (2) personalized topic feeds and revision history; (3) cross-edition trend analysis and misinformation detection; (4) mobile application development; and (5) domain-specific fine-tuning of the sentiment model on Indian regional news corpora.

VI. ACKNOWLEDGMENT

The authors gratefully acknowledge the Department of Computer Science and Engineering, Barkatullah University Institute of Technology (BUIT), Bhopal, for providing the computational resources and institutional support necessary for this research. The authors also thank the student volunteers who participated in the usability evaluation study.

REFERENCES

[1] B. Singh and A. Sharma, "Current Affairs Preparation Patterns Among Competitive Exam Aspirants in India: A Survey," Journal of Educational Research and Practice, vol. 12, no. 3, pp. 45-58, 2023.
[2] R. Smith, "An Overview of the Tesseract OCR Engine," Proc. 9th Int. Conf. on Document Analysis and Recognition (ICDAR), pp. 629-633, 2007.
[3] Y. Xu et al., "LayoutLM: Pre-training of Text and Layout for Document Image Understanding," Proc. ACM SIGKDD, pp. 1192-1200, 2020.
[4] Google Cloud, "Document AI - Intelligent Document Processing," [Online]. Available: https://cloud.google.com/document-ai, 2024.
[5] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for NLG," Proc. ACL, pp. 7871-7880, 2020.
[6] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," JMLR, vol. 21, no. 140, pp. 1-67, 2020.
[7] M. Grusky, M. Naaman, and Y. Artzi, "Newsroom: A Dataset of 1.3 Million Summaries," Proc. NAACL-HLT, pp. 708-719, 2018.
[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers," Proc. NAACL-HLT, pp. 4171-4186, 2019.
[9] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, Morgan & Claypool, 2012.
[10] R. Nallapati et al., "Abstractive Text Summarization Using Sequence-to-Sequence RNNs," Proc. CoNLL, pp. 280-290, 2016.
[11] A. Islam, S. Akter, and M. Hossain, "An Adaptive Learning DSS for Current Affairs," Expert Systems with Applications, vol. 189, 2022.
[12] Streamlit Documentation, Streamlit Apps and UI Development, [Online]. Available: https://docs.streamlit.io/
[13] Meta AI, LLaMA 3.2 Vision - Multimodal LLM Overview, [Online]. Available: https://ai.meta.com/llama/
[14] C. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," Proc. ACL Workshop on Text Summarization Branches Out, pp. 74-81, 2004.
[15] T. Zhang et al., "BERTScore: Evaluating Text Generation with BERT," Proc. ICLR, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)