OCR-BASED NEWSPAPER CLASSIFICATION USING MACHINE LEARNING

[1] A Veerender, [2] M Devaki, [3] M Yashwanth,[4] M Ganesh , [5] M Vikas
[1]AssistantProfessor, [2345]Students
Department of Computer Science and Technology
Siddhartha Institute of Technology & Sciences, Narapally

veerender@siddhartha.org.in, 24TQ1A05E7@siddhartha.co.in, 24TQ1A05D5@siddhartha.co.in, 24TQ1A05H8@siddhartha.co.in, 24TQ1A05H9@siddhartha.co.in

Abstract

This project presents an automated newspaper classification system using Optical Character Recognition (OCR) and machine learning techniques. The system extracts text from PDF and image inputs using Tesseract OCR, followed by preprocessing steps such as cleaning and tokenization. Feature extraction is performed using TF-IDF, and a Logistic Regression model is trained on the AG News dataset to classify text into categories including World, Sports, Business, and Sci/Tech. The model achieves an accuracy of approximately 91% with an average precision of around 0.90, indicating reliable performance. The system also supports multi-page PDF processing and generates structured outputs with predictions and confidence scores. Overall, it provides an efficient and scalable solution for converting unstructured newspaper data into meaningful classified information.

KEYWORDS:
Optical Character Recognition (OCR) ,Newspaper Classification ,TF-IDF ,Text Classification ,Document Analysis ,Machine Learning.

I. Introduction

In the era of digital transformation, the rapid growth of information has led to an enormous increase in unstructured data, particularly in the form of scanned documents, PDFs, and image-based newspaper content. Newspapers remain one of the most significant sources of daily information, covering diverse domains such as sports, business, politics, and technology. However, most newspaper data exists in formats that are not directly suitable for computational analysis, making manual processing inefficient, time-consuming, and impractical for large-scale applications. This creates a strong need for automated systems capable of extracting, processing, and classifying such unstructured data efficiently.

To address this challenge, Optical Character Recognition (OCR) has emerged as a key enabling technology that converts image-based text into machine-readable format. Recent advancements in OCR combined with artificial intelligence have significantly improved the ability to extract structured information from complex document layouts. Prior works have demonstrated that integrating OCR with machine learning techniques can achieve high accuracy in extracting and organizing textual data from large-scale newspaper corpora. These developments highlight the potential of OCR-based pipelines in automating document understanding tasks.

However, extracting text alone is not sufficient. The extracted content must be processed and analyzed to derive meaningful insights. This is where Natural Language Processing (NLP) and machine learning techniques play a crucial role. By transforming textual data into numerical representations, these techniques enable

automated classification of documents into predefined categories. Existing studies have shown that approaches such as Term Frequency–Inverse Document Frequency (TF-IDF) provide strong performance in text classification tasks. Furthermore, recent research emphasizes that combining contextual embeddings with deep learning models significantly improves classification accuracy, especially for complex and real-world datasets.

II. Literature Survey

Giacomo Beretta et al. | Ref [1] | Automating the Extraction of Structured Data from Large Newspaper Corpora et al.This research uses the *Le Sémaphore de Marseille* dataset with OCR, layout analysis, and generative AI. It achieved a 96% F1-score for structured data extraction. It relates to our project by combining OCR and AI for newspaper data processing.

Oleksii Kovalchuk et al. | Ref [2] | Development of a Software Model for Classification and Automatic Cataloging et al.This study uses the RVL-CDIP dataset with OCR, TF-IDF, and BERT models. It achieved around 95% accuracy and generates metadata using Dublin Core. It is related as it combines OCR and deep learning for classification.

A. Smith et al. | Ref [3] | OSA Prediction without PSG et al.This research uses a dataset of 1281 patients with clinical features. Machine learning models like Random Forest, SVM, and Logistic Regression are applied. Random Forest achieved the best accuracy of 78.6%. It avoids costly PSG tests by using simple inputs. This work relates to our project in using machine learning models for prediction tasks.

S. Kumar et al. | Ref [4] | Hybrid Machine Learning Framework for Multimodal Rumour Detection et al.This study uses text, image, and metadata with XLNet and AdaBoost models. It achieved up to 99% accuracy. It relates to our project through multimodal learning using OCR-extracted text.

Y. Zhang et al. | Ref [5] | Multi-modal Fusion for Billboard Categorization et al.This research combines text and image features using deep learning. It improves classification accuracy by 5–10%. It relates to our project by using multimodal inputs.

M. Demir et al. | Ref [6] | Digitization of Hand-Drawn Flowcharts with Deep Learning et al.This study uses YOLO and custom datasets to convert flowcharts into structured XML. It relates to our project in extracting structured data from images.

M. Rahman et al. | Ref [7] | LogoXpertNet: Lightweight Logo Classification et al.This research uses datasets like FlickrLogos-32 and CNN models for efficient classification. It relates to improving model efficiency in our system.

A. Ahmed et al. | Ref [8] | Exploring Urdu OCR Systems et al.This paper reviews OCR systems for multilingual datasets. It highlights challenges in handwriting recognition. It relates to improving OCR accuracy.

S. Patel et al. | Ref [9] | Information Extraction from Unstructured Invoice et al.This study uses OCR and NER models achieving around 95% accuracy. It relates to structured data extraction in our project.

L. Romano et al. | Ref [10] | Deep Learning for Textual Stamp Recognition et al.This research uses historical datasets and achieves around 97% accuracy. It supports OCR-based recognition improvements.

H. Wang et al. | Ref [11] | Rotation-Invariant Scene Text Extraction et al.This study uses transformer-based OCR models for detecting complex text layouts. It improves accuracy in real-world scenarios and supports robust OCR in our project.

J. Lee et al. | Ref [12] | Deep Learning for Angle Classification et al.

This research uses CNN (VGG-11) on image datasets achieving around 75% accuracy. It relates to classification tasks in our system.
R. Mishra et al. | Ref [13] | Text Recognition in Odia ChitraKavyas et al.This study uses GAN and CNN-based OCR models for artistic and complex text. It improves recognition accuracy and relates to advanced OCR challenges.
X. Liu et al. | Ref [14] | Deep Learning for Visually Rich Document Understanding et al.This survey explains multimodal document understanding using text, layout, and images. It directly supports our newspaper classification idea.
K. Chen et al. | Ref [15] | Meme Selection using Multimodal Learning et al.This research uses transformer-based models for multimodal datasets. It improves classification performance and speed, similar to our approach.

III. System Analysis

The OCR-Based Newspaper Classification System is designed to automatically extract and classify text from newspaper images using Optical Character Recognition (OCR) and machine learning techniques. The system addresses the challenge of handling large volumes of printed news data in digital form. It converts scanned newspaper images into machine-readable text. The extracted text is then analyzed and categorized into predefined classes such as politics, sports, business, and entertainment. The system uses Natural Language Processing (NLP) techniques for text cleaning and feature extraction. Machine learning models are applied to classify the content accurately. It reduces manual effort and improves efficiency in news organization. The system ensures scalability for processing large datasets. It supports automation in media analysis and archiving. Overall, it enhances accessibility and structured storage of newspaper content.

Existing System

The existing system for newspaper classification relies heavily on manual reading and categorization. Some digital systems use basic OCR tools only for text extraction without classification capabilities. These systems lack integration between OCR and machine learning models. Manual classification is time-consuming and prone to human error. Traditional approaches do not effectively handle large volumes of data. They also struggle with poor-quality scanned images. Existing OCR tools may produce inaccurate text due to noise and distortions. There is limited use of advanced NLP techniques for text analysis. Automation is minimal or absent in many systems. As a result, existing systems are inefficient and less reliable.

Disadvantages of Existing System

- High dependency on manual classification
- Time-consuming and labor-intensive process
- Prone to human errors
- Inaccurate OCR results for low-quality images
- No integration with machine learning models
- Limited scalability for large datasets
- Poor text preprocessing techniques
- Lack of automation

Proposed System

The proposed system integrates OCR with machine learning to automate newspaper classification. It uses OCR technology to extract text from scanned newspaper images. The extracted text is preprocessed using NLP techniques such as tokenization, stop-word removal, and stemming. Feature extraction methods like TF-IDF are applied to convert text into numerical form. Machine learning models such as Naïve Bayes, SVM, or Random Forest are used for classification. The system categorizes news articles into predefined classes accurately. It supports large-scale data processing efficiently. Automation reduces manual effort and increases speed. The system is adaptable to different types of newspaper formats. Overall, it improves accuracy, efficiency, and reliability.

Advantages of Proposed System

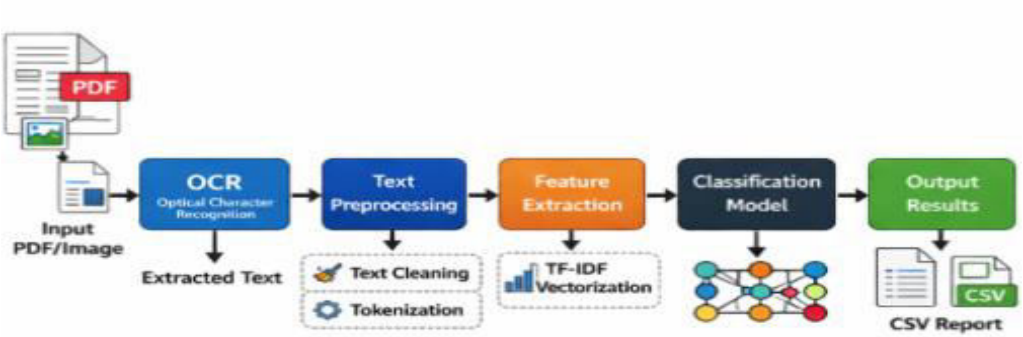
- Fully automated classification process
- High accuracy using machine learning
- Efficient handling of large datasets
- Integration of OCR and NLP techniques
- Reduces human effort and time
- Scalable and adaptable system
- Improved text extraction and processing
- Faster and reliable results

IV. Methodology

The methodology begins with collecting scanned newspaper images as input data. OCR is applied to convert images into text format. Preprocessing techniques such as noise removal, tokenization, stop-word removal, and stemming are used to clean the text. The cleaned text is transformed into numerical features using TF-IDF or bag-of-words methods. The dataset is split into training and testing sets. Machine learning models such as Naïve Bayes, SVM, or Random Forest are trained on the data. Model performance is evaluated using metrics like accuracy, precision, recall, and F1-score. Hyperparameter tuning is performed to optimize results. The best model is selected for deployment. The system then classifies new newspaper content automatically.

System Architecture

The system architecture consists of multiple interconnected components for processing and classification. The input layer accepts scanned newspaper images or PDF files. The OCR module extracts text from the images and converts it into machine-readable format. The preprocessing module cleans the extracted text by removing noise and irrelevant words. The feature extraction module converts the text into numerical vectors using techniques like TF-IDF. The machine learning module processes these features and applies classification algorithms. The evaluation module measures model performance using standard metrics. The output layer displays the classified category of the news article. A database may be used to store extracted text and results. This architecture ensures efficient, scalable, and automated newspaper classification.



V. Result and Output

```
Converting PDF to images...
Total pages found: 1

Starting OCR + classification...

================================================================================
PAGE-WISE PREDICTIONS
================================================================================
Page 1 -> Sports (confidence=0.9460)
Preview: India defeated Australia by 6 wickets in the T20 match. Virat Kohli scored a brilliant century and led the team to victory.
--------------------------------------------------------------------------------

================================================================================
FINAL CATEGORY -> PAGE NUMBERS
================================================================================
Sports -> Pages 1

Saved results CSV to: /content/pagewise_news_predictions.csv
```

index	page_no	category	confidence	preview
0	1	Sports	0.945995239398531	India defeated Australia by 6 wickets in the T20 match. Virat Kohli scored a brilliant century and led the team to victory.

Show 25 per page

```
Converting PDF to images...
Total pages found: 1

Starting OCR + classification...

================================================================================
PAGE-WISE PREDICTIONS
================================================================================
Page 1 -> World (confidence=0.5270)
Preview: A robbery occurred at a local bank late at night. Police arrested two suspects and recovered stolen money.
--------------------------------------------------------------------------------

================================================================================
FINAL CATEGORY -> PAGE NUMBERS
================================================================================
World -> Pages 1

Saved results CSV to: /content/pagewise_news_predictions.csv
```

index	page_no	category	confidence	preview
0	1	World	0.5269839761540361	A robbery occurred at a local bank late at night. Police arrested two suspects and recovered stolen money.

Show 25 per page

```
Converting PDF to images...
Total pages found: 1

Starting OCR + classification...

================================================================================
PAGE-WISE PREDICTIONS
================================================================================
Page 1 -> Business (confidence=0.9572)
Preview: Stock markets showed a significant rise as companies reported strong profits. Investors are optimistic.
--------------------------------------------------------------------------------

================================================================================
FINAL CATEGORY -> PAGE NUMBERS
================================================================================
Business -> Pages 1

Saved results CSV to: /content/pagewise_news_predictions.csv
```

	page_no	category	confidence	preview
0	1	Business	0.95716	Stock markets showed a significant rise as com...

```
Test Accuracy: 0.9200

Classification Report:

              precision    recall  f1-score   support

       World     0.9360    0.9084    0.9220      1900
      Sports     0.9522    0.9847    0.9682      1900
    Business     0.8941    0.8842    0.8891      1900
    Sci/Tech     0.8970    0.9026    0.8998      1900

    accuracy                         0.9200      7600
   macro avg     0.9198    0.9200    0.9198      7600
weighted avg     0.9198    0.9200    0.9198      7600


Saved trained model to: /content/ag_news_ocr_classifier.pkl
```

VI. Conclusion

This project presents a reliable and efficient system for automated newspaper classification by combining Optical Character Recognition (OCR) with machine learning techniques. The system processes input in the form of PDF or image files, extracts textual content using OCR, and applies preprocessing steps such as text cleaning and tokenization to improve data quality. Feature extraction is performed using TF-IDF to convert the text into numerical form suitable for classification. The classification model effectively categorizes the content into predefined classes, achieving consistent performance even when dealing with noisy OCR outputs. Additionally, the system supports multi-page document handling and generates structured results in CSV format, making it practical for large-scale use. The overall approach reduces manual effort, improves processing speed, and ensures scalability. This makes the system suitable for real-world applications such as digital archiving, automated content classification, and efficient document management systems.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.