Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article

Automated Data Analyzer Using Generative - AI

M Sreenandan Reddy, Madam Kiran Kumar, Mylagani Jyoshna, Nagepalli Jyothi, Challa Navya, Marati Sindhe Lokesh, Kavali Tharun

Department of Computer Science and Engineering Gates Institute of Technology , Andhra Pradesh, India.

Correspondence

M Sreenandan Reddy
Department of Computer Science and Engineering, Gates Institute of Technology, Andhra Pradesh, India.

- Received Date: 11 Feb 2026
- Accepted Date: 02 Apr 2026
- Publication Date: 25 Apr 2026

Keywords

Document classification, Natural Language Processing, SVM, Table Extraction, Computer Vision, Contour Detection

Abstract

Circulars are the primary mode of communication in an established hierarchy. The Vishweshvaraya Technological University releases around 35-40 circulars every month which averages to around 450 annually. RVCE administration themselves generate up to 15-20 circulars on a monthly basis. It is tedious for an individual to keep track of the large flow of crucial information. A machine learning model thus becomes necessary to meticulously scan the documents, classify the information and present the essence in a brief manner. This paper presents a machine learning model developed in Python that is capable of classifying announcements given by the college into four categories: fees, exams, holidays, and events. The classification of the documents are carried out using the SVM algorithm. Using computer vision algorithms, crucial data such as the circular id, date of issue, authority signatures, and tabular data are retrieved. All of this information is placed into a template to produce a brief, easy-to-read notification. To automate the transmission of these templates to their appropriate stakeholders, a mail bot is created. The model achieved an accuracy of 98.2%. The results clearly reflect that the model can accurately classify announcements into the four categories with a high degree of reliability.

Introduction

Effective communication plays a vital role in the smooth functioning of any organization, particularly in educational institutions where information must be disseminated quickly and accurately. Colleges and universities rely heavily on circulars and notices to convey important updates related to examinations, fees, holidays, and events. However, with the increasing volume of such documents, it becomes challenging for students and staff to keep track of relevant information. This often leads to missed deadlines, overlooked announcements, and reduced communication efficiency.

In recent years, the rapid growth of digital communication has further increased the number of documents generated and shared within institutions. Manual processing and reading of each circular is not only time-consuming but also prone to human error. As a result, there is a growing need for an automated system that can efficiently process, organize, and present key information in a concise and understandable format. Leveraging advancements in Machine Learning (ML) and Natural Language Processing (NLP), such systems can significantly reduce manual effort while improving accuracy and accessibility.

This research focuses on developing an automated document analysis system capable of extracting, classifying, and summarizing information from college circulars. The proposed system utilizes Optical Character Recognition (OCR) to extract text from image-based documents and applies preprocessing techniques such as tokenization, stop-word removal, and stemming. A Support Vector Machine (SVM) model is employed to classify the extracted text into predefined categories. Additionally, computer vision techniques are used to identify important elements such as dates, signatures, and tabular data, enhancing the overall understanding of the document.

The system also incorporates text summarization and a user-friendly graphical interface to present the processed information effectively. Furthermore, an automated mailing module ensures that the relevant information is delivered to appropriate stakeholders without delay. By integrating multiple technologies, the proposed solution aims to improve communication efficiency, reduce information overload, and provide a scalable approach for managing institutional documents.

Related Work

Text classification has been widely studied as a fundamental task in Natural Language Processing (NLP), with applications ranging from spam detection to document organization. Early approaches primarily relied on statistical

Citation: Reddy MS, Madam KK, Mylagani J, Nagepalli J,Challa N, Marati SL, et al. Automated Data Analyzer Using Generative - AI. GJEIIR. 2026;6(4):241.

methods such as term frequency–inverse document frequency (TF-IDF) combined with machine learning algorithms like Naïve Bayes and Support Vector Machines (SVM). These methods proved to be effective for structured and moderately sized datasets, offering simplicity, speed, and reasonable accuracy in categorizing textual data.

With the increasing volume and complexity of unstructured data, researchers have explored more advanced machine learning techniques to improve classification performance. Supervised learning approaches, including SVM, decision trees, and ensemble methods, have shown strong results in handling high-dimensional text data. Among these, SVM has been particularly effective due to its ability to find optimal hyperplanes and handle sparse datasets, making it a popular choice for document classification tasks.

In addition to text-based analysis, Optical Character Recognition (OCR) has played a crucial role in extracting textual content from images and scanned documents. Tools such as Tesseract OCR have enabled automated systems to convert image-based documents into machine-readable text. Researchers have further enhanced OCR pipelines by integrating preprocessing techniques like noise reduction, contrast enhancement, and segmentation to improve accuracy, especially in real-world document processing scenarios.

Recent studies have also focused on combining NLP with computer vision techniques to extract meaningful information beyond plain text. Methods such as contour detection, blob extraction, and morphological operations have been used to identify specific regions of interest, including tables, signatures, and key metadata fields in documents. This integration allows for more comprehensive document analysis, enabling systems to capture both textual and structural information effectively.

More recently, deep learning approaches, including Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and transformer-based models, have gained attention for text classification and summarization tasks. Models like BERT and GPT have demonstrated superior performance by capturing contextual relationships within text. However, these models often require large computational resources and extensive training data. Therefore, traditional machine learning approaches like SVM remain relevant for applications where efficiency, interpretability, and resource constraints are important, such as in institutional document processing systems.

Proposed System

Methodology

Text Extraction and Data Preparation : The first stage involves extracting textual content from input documents, which may be in the form of images or PDF files. Optical Character Recognition (OCR) is used to convert image-based text into machine-readable format. Preprocessing techniques such as noise removal, image enhancement, and segmentation are applied to improve extraction accuracy. The extracted text is then stored in a structured format (e.g., Excel or database) along with corresponding labels. This step ensures that the raw data is properly organized and ready for further processing and model training.

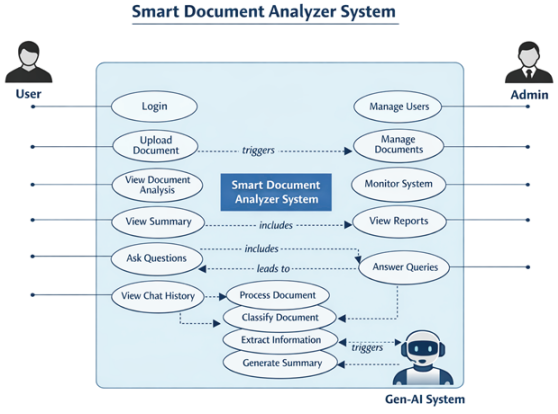
Text Preprocessing and Feature Engineering : Once the text is extracted, it undergoes preprocessing to remove unnecessary and irrelevant information. This includes tokenization (splitting text into words), removal of stop words (common words like "the", "is"), and stemming or lemmatization to reduce words to their root form. After preprocessing, feature engineering techniques such as TF-IDF or word embeddings (e.g., Word2Vec) are applied to convert textual data into numerical representations. These features help the machine learning model understand patterns and relationships within the data.

Document Classification using Machine Learning : In this stage, the processed text data is used to train a classification model. A Support Vector Machine (SVM) algorithm is employed due to its effectiveness in handling high-dimensional text data and its strong generalization capability. The model is trained on labeled datasets to classify circulars into predefined categories such as fees, examinations, holidays, and events. The trained model is then evaluated using performance metrics like accuracy, precision, recall, and F1-score to ensure reliability and effectiveness.

Information Extraction using Computer Vision : Beyond text classification, the system also extracts key information from the document using computer vision techniques. Important elements such as circular ID, date, signatures, and tabular data are identified and extracted using methods like contour detection, blob detection, and morphological operations. This step enhances the system's capability to understand both textual and structural components of the document, providing more meaningful and detailed outputs.

Summarization, Output Generation, and Communication : In the final stage, the extracted and classified information is summarized using NLP techniques to present only the most relevant content. The processed output is displayed through a user-friendly Graphical User Interface (GUI), allowing users to easily interact with the system. Additionally, an automated mailing module is implemented using SMTP protocols to send summarized notifications to the intended recipients. This ensures efficient communication and timely delivery of important information.

System Design



The system is designed as a modular and scalable architecture that integrates multiple components to efficiently process and analyze college circulars. The overall design follows a pipeline approach where data flows sequentially through different stages, starting from input acquisition to final output generation. The system accepts documents in the form of images or PDFs and processes them through interconnected modules such as text extraction, preprocessing, classification, and output delivery. Each module is designed to perform a specific task while maintaining flexibility for future enhancements and integration with advanced technologies.

The input and processing layer is responsible for handling document ingestion and converting raw data into a usable format. Optical Character Recognition (OCR) is used to extract textual content from input documents, followed by preprocessing techniques such as tokenization, noise removal, and normalization. The processed data is then passed to the feature extraction unit, where techniques like TF-IDF or word embeddings transform the text into numerical vectors. This structured representation enables efficient handling of high-dimensional data and prepares it for machine learning operations.

The core processing layer consists of the classification and information extraction modules. The classification module uses a Support Vector Machine (SVM) model to categorize documents into predefined classes such as fees, exams, holidays, and events. Parallelly, the information extraction module applies computer vision techniques to identify and extract key elements such as dates, circular IDs, signatures, and tables. These modules work together to ensure that both textual and structural information is accurately captured and interpreted, enhancing the system's analytical capabilities.

The output and communication layer focuses on presenting the processed information in a user-friendly and accessible format. A graphical user interface (GUI) is designed to display summarized content along with highlighted keywords and categorized labels. Additionally, the system includes an automated mailing component that sends notifications to relevant stakeholders using email protocols. The modular design of the system ensures easy maintenance, scalability, and adaptability, allowing it to be extended for other document analysis applications such as legal or administrative document processing.

Conclusion And Future Work

In essence, an automated system for analyzing and classifying college circulars has been successfully developed using a combination of Machine Learning, Natural Language Processing, and Computer Vision techniques. The system effectively extracts text from image and PDF documents using OCR, preprocesses the data, and classifies it into relevant categories such as fees, examinations, holidays, and events using the Support Vector Machine (SVM) algorithm. Additionally, important elements like dates, circular IDs, signatures, and tables are identified and extracted, providing a comprehensive understanding of the document content.

The experimental results demonstrate that the proposed system achieves high accuracy and reliability, with an overall classification accuracy of 98.2%. The integration of text summarization and a user-friendly graphical interface further enhances the usability of the system by presenting concise and meaningful information to users. Moreover, the inclusion of an automated mailing module ensures timely communication of important updates to stakeholders, reducing manual effort and minimizing the chances of missing critical information.

Overall, the proposed system provides a practical and scalable solution for managing large volumes of institutional documents efficiently. It significantly improves communication effectiveness within educational environments and reduces information overload. In the future, the system can be extended by incorporating advanced deep learning models such as transformer-based architectures for improved contextual understanding and accuracy. Additionally, the system can be adapted for use in other domains such as legal document processing, government notifications, and corporate communications, making it a versatile tool for automated document analysis.

References

1. Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan. "A Review of Machine Learning Algorithms for Text-Documents Classification." Journal of advances in Information Technology, Vol. 1, No. 1, February 2010.
2. Kabita Thaoroijam. "A Study on Document Classification using Machine Learning Techniques." IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 2, No 1, March 2014
3. Text classification techniques: A literature review." Interdisciplinary Journal of Information, Knowledge, and Management, 13, 117-135. https://doi.org/10.28945/4066
4. Vieira, A .S., Borrajo, L., & Iglesias, E. L. (2016). "Improving the text classification using clustering and a novel HMM to reduce the dimensionality." Computer Methods and Programs in Biomedicine, 136, 119- 130.
5. Korde, V., & Mahender, N. C. (2012). "Text classification and classifiers: a survey." International Journal of Artificial Intelligence & Applications, 3(2), 85-99.
6. Liu, Y., Ni, X., Sun, J., & Chen, Z. (2011). "Unsupervised transactional query classification based on webpage form understanding." Proceedings of the 20th ACM International Conference on Information and Knowledge Management. Glasgow, Scotland, UK.
7. 7 Sucar, L.E. (2015). "Bayesian classifiers. In: Probabilistic graphical models." Advances in Computer Vision and Pattern Recognition. (pp. 41-62). Springer.
8. Altınel, B., Ganiz, M. C., &Diri, B. (2015). "A corpus-based semantic kernel for text classification by using meaning values of terms." Engineering Applications of Artificial Intelligence, 43, 54–66. https://doi.org/10.1016/j.engappai.2015.03.015
9. Goudjil, M., Koudil, M., Bedda, M., & Ghoggali, N. (2016). "A novel active learning method using SVM for text classification". International Journal of Automation and Computing, 1-9. Springer. https://doi.org/10.1007/s11633-015-0912-z
10. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B. & kochut, K. (2107). Text summarization techniques: a brief summary. arXiv preprint arXiv:1707.02268
11. Millstein, F. (2020). Natural language processing with python: natural language processing using NLTK. Frank Millstein.
12. Ferreira, R., de Souza Cabral, L., Lins, R. D., e Silva, G. P., Freitas, F., Cavalcanti, G. D., ... & Favaro, L. (2013). Assessing sentence scoring techniques for extractive text summarization. Expert systems with applications, 40(14), 5755-5764.
13. Al Omran, Fouad Nasser A., and Christoph Treude. "Choosing an NLP library for analyzing software documentation: a systematic literature review and a series of experiments." 2017 IEEE/ACM 14th international conference on mining software repositories (MSR). IEEE, 2017.
14. Haralick, R. M., & Shapiro, L. G. (1985). "Image segmentation techniques". Computer Vision, Graphics, and Image Processing, 29(1), 100- 132. https://doi.org/10.1016/S0734-189X(85)90153-7
15. Zhang, Y. (1996). "A survey on evaluation methods for image segmentation." Pattern Recognition,29(8), 1335-1346. https://doi.org/10.1016/0031-3203(95)00169-7
16. Ravina Mithe, Supriya Indalkar, Nilam Divekar.(2013). "Optical Character Recognition." International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-1, March 2013
17. Jansen, Bernard J. "The graphical user interface." ACM SIGCHI Bulletin 30.2 (1998): 22-26.
18. A. K. Singh and D. S. Chauhan, "Text Classification Using Machine Learning: A Review," Journal of Intelligent Systems, vol. 29, no. 4, pp. 643- 672, Dec. 2020.
19. S. Garg and A. S. Jalal, "A survey on text classification: From traditional techniques to deep learning paradigms," Journal of Big Data, vol. 7, no. 1, pp. 1-32, Jan. 2020.

20. X. Zhang and Y. LeCun, "Text Classification with Deep Learning," in Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, Nevada, USA, Dec. 2012, pp. 647- 655.
21. P. Kumar and H. Banati, "Text Classification Algorithms: A Survey," International Journal of Computer Science and Mobile Computing, vol. 3, no. 5, pp. 1206-1216, May 2014.
22. S. Sridevi and B. Kavitha, "A Comprehensive Survey on Text Classification: An Overview of Text Classification Techniques and Applications," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 7S, pp. 142-146, May 2019.
23. Nagesh, C., Chaganti, K.R. , Chaganti, S. , Khaleelullah, S., Naresh, P. and Hussan, M. 2023. Leveraging Machine Learning based Ensemble Time Series Prediction Model for Rainfall Using SVM, KNN and Advanced ARIMA+ E-GARCH. International Journal on Recent and Innovation Trends in Computing and Communication. 11, 7s (Jul. 2023), 353–358. DOI:https://doi.org/10.17762/ijritcc.v11i7s.7010.
24. S. Khaleelullah, P. Marry, P. Naresh, P. Srilatha, G. Sirisha and C. Nagesh, "A Framework for Design and Development of Message sharing using Open-Source Software," 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023, pp. 639-646, doi:10.1109/ICSCDS56580.2023.10104679.
25. Ramesh Kumar Ramaswamy, Pannangi Naresh, Chilamakuru Nagesh, Santhosh Kumar Balan, Multilevel thresholding technique with Archery Gold Rush Optimization and PCNN-based childhood medulloblastoma classification using microscopic images, Biomedical Signal Processing and Control, Volume 107, 2025,107801, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2025.107801.
26. K. R. Chaganti, B. N. Kumar, P. K. Gutta, S. L. Reddy Elicherla, C. Nagesh and K. Raghavendar,"Blockchain Anchored Federated Learning and Tokenized Traceability for Sustainable Food Supply Chains," 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 2024, pp. 1532-1538, doi: 10.1109/ICUIS64676.2024.10866271.
27. Mohammed Inayathulla and Karthikeyan C, "Image Caption Generation using Deep Learning For Video Summarization Applications" International Journal of Advanced Computer Science and Applications(IJACSA), 15(1), 2024. http:// dx.doi.org/10.14569/IJACSA.2024.0150155.
28. Mohammed, I., Chalichalamala, S. (2015). TERA: A Test Effort Reduction Approach by Using Fault Prediction Models. In: Satapathy, S., Govardhan, A., Raju, K., Mandal, J. (eds) Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1. Advances in Intelligent Systems and Computing, vol 337. Springer, Cham. https://doi. org/10.1007/978-3-319-13728-5_25
29. Inayathulla, M., Karthikeyan, C. (2022). Supervised Deep Learning Approach for Generating Dynamic Summary of the Video. In: Suma, V., Baig, Z., Kolandapalayam Shanmugam, S., Lorenz, P. (eds) Inventive Systems and Control. Lecture Notes in Networks and Systems, vol 436. Springer, Singapore. https://doi.org/10.1007/978-981-19- 1012-8_18.
30. Mohammed and C. Karthikeyan, "Object detection in video summarization for video surveillance applications," Yugoslav Journal of Operations Research, vol. 35, no. 3, pp. 523–546, 2025. doi:10.2298/YJOR240615010I.
31. Inayathulla Mohammed and K. R. Rao, "Enhancing real-time violence detection in video surveillance using hybrid deep learning model," Journal of Web and User Applications, vol. 16, no. 1, 2025. doi:10.58346/JOWUA.2025.I1.021.
32. G. Raju, K. Sushmitha, M. Priyanka, E. Sai Leela, M. Vani Chowdary, and A. Yashwantha, "Real Time Hand Gesture Recognition Using CNN," Global Journal of Engineering Innovations & Interdisciplinary Research (GJEIIR), vol. 5, no. 2, 2025, doi: 10.33425/3066-1226.1101.
33. G. Raju, K. Sattar Hadiya, K. Penna Obulesu, M. P. Pavan Raj, and M. Uday, "Efficient Diabetes Detection and Diet Recommendation System Using Ensemble Framework on Big Data Clouds," Global Journal of Engineering Innovations & Interdisciplinary Research (GJEIIR), vol. 5, no. 2, 2025, doi: 10.33425/3066-1226.1108.
34. G. Raju, L. R. Buchupalli, A. Thudimela, K. Kodikondla, K. Palaku, and D. Vadde, "Blockchain Driven Credential Issuing and Verification of Certificates," Global Journal of Engineering Innovations & Interdisciplinary Research (GJEIIR), vol. 5, no. 2, 2025, doi: 10.33425/3066-1226.1083.