WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Automated Document Analysis

Shubham S Manachekar[1], Aditya P Pyati[2], Amey A Sangle[3], Madhavi A More[4], Babychen Mathew[5]

[1] Electronics & Computer Science Shah & Anchor Kutchhi Engineering College Mumbai, India shubham.manachekar16226@sakec.ac.in
[2] Electronics & Computer Science Shah & Anchor Kutchhi Engineering College Mumbai, India aditya.pyati16209@sakec.ac.in
[3] Electronics & Computer Science Shah & Anchor Kutchhi Engineering College Mumbai, India amey.sangle16267@sakec.ac.in
[4] Electronics & Computer Science Shah & Anchor Kutchhi Engineering College Mumbai, India madhavi.more16308@sakec.ac.in
[5] Associate Professor , Electronics & Computer Science Shah & Anchor Kutchhi Engineering College Mumbai, India babychen.mathew@sakec.ac.in

ABSTRACT—

This project presents a Document Analysis Application using advanced technologies. It includes OCR, Image Caption Generation, Document Classification, and Spacy-based Text Summarization. OCR by Tesseract and EasyOCR extracts text accurately from diverse sources. Image Caption Generation uses Vision Studio Azure and deep learning for better image understanding. Document Classification categorizes documents into domains with SVM and Multinomial Naive Bayes classifiers. Text Summarization, powered by Spacy, effectively condenses documents. Future improvements target classification models, language support, data source integration, and user interfaces. The application offers efficient document analysis for decision-making in the digital era.

Keywords—. Document Analysis, Machine learning, OCR, Image Captioning, Document classification, Summarization, NLP

Introduction :

Automated Document Analysis (ADA) has become a crucial area within the field of information technology, facilitating the efficient processing and examination of textual documents in various sectors. Recent advancements in incorporating sophisticated methods like Optical Character Recognition (OCR), document categorization, and extractive summarization have transformed the landscape of document analysis, empowering entities to extract valuable insights from unstructured textual data.

The utilization of OCR technology, particularly through tools such as Tesseract, has greatly improved the functionalities of ADA by allowing the retrieval of text from a wide range of document formats, including PDF and Word files. This procedure involves converting scanned documents or images into text that can be read by machines (txt format), thereby enabling further examination and processing.

Furthermore, ADA has expanded its scope to include image processing capabilities alongside textual analysis. In instances where documents incorporate images, ADA systems are equipped with image capturing and captioning features to derive meaningful information from visual content. This integration enables a comprehensive analysis of documents containing both text and visual elements, enhancing the depth and breadth of insights obtained from document analysis.

Document categorization is another essential aspect of ADA, enabling the classification of documents into predetermined classes or categories. Within engineering fields, document categorization is crucial for organizing and managing technical documents effectively. By utilizing machine learning techniques like the Naive Bayes theorem, ADA systems can categorize documents into various engineering branches such as computer science, electronics, cyber security, VLSI, and more. This categorization process facilitates the efficient retrieval and management of domain-specific documents, thereby streamlining workflows and decision-making processes.

Additionally, the incorporation of extractive summarization with tools like Spacy further enhances the analytical capabilities of ADA systems. Through training models to recognize and extract key sentences or phrases from documents, extractive summarization enables the creation of concise summaries while preserving the essence and crucial information present in the original text. The adaptability of extractive summarization allows for the generation of summaries in different lengths, catering to the specific needs and preferences of users.

Overall, the combination of OCR, image processing, document categorization, and extractive summarization techniques in ADA systems signifies a transformative shift in document analysis. By leveraging these advanced capabilities, organizations can extract actionable insights, enhance decision-making processes, and uncover new avenues for innovation across various domains.

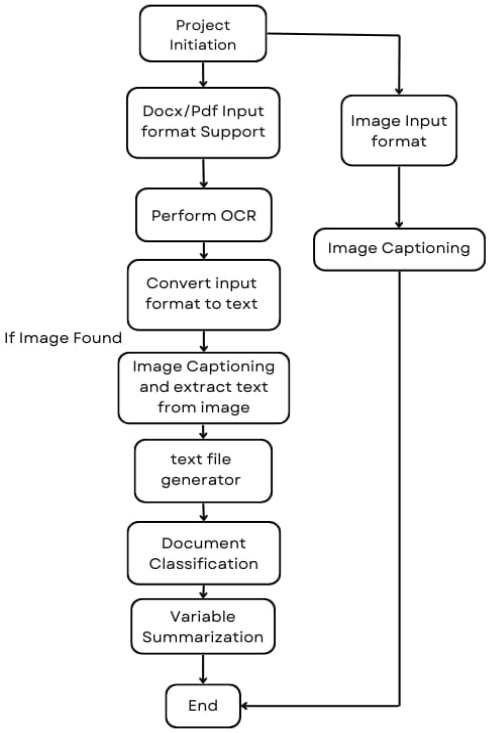


fig : basic flowchart of the designed model

Methodologies and Techniques

Automated Document Analysis (ADA) involves a comprehensive approach that combines various methodologies and techniques to derive valuable insights from textual documents. This section explores the methodologies and techniques utilized in four primary modules of ADA: Optical Character Recognition (OCR), Image Processing, Document Classification, and Extractive Summarization.

1. Optical Character Recognition (OCR) Module

The OCR module acts as the foundational element of ADA, enabling the transformation of scanned documents or images with text into machine-readable formats like plain text (txt). Tesseract OCR, a popular open-source OCR engine, is utilized for its reliability and precision in extracting text from various document formats like PDF and Word files. The OCR process comprises several stages:

Preprocessing: Before OCR, the document undergoes preprocessing steps such as noise reduction, image enhancement, and binarization to enhance the input image's quality.

Text Extraction: Tesseract employs advanced algorithms to identify and extract text regions from the input image while preserving the original document's layout and formatting.

Conversion to Txt Format: The extracted text is transformed into plain text (txt) format, enabling further analysis and processing.

2. Image Processing Module

In instances where documents feature images, the Image Processing module plays a vital role in extracting meaningful data from visual content. This module consists of two key functions: image capturing and image captioning.

Image Capturing: ADA systems utilize image capturing techniques to retrieve images embedded within documents. This process may involve analyzing the document structure to identify image elements and extracting them for subsequent analysis.

Image Captioning: By utilizing deep learning models like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), image captioning algorithms generate descriptive captions or labels for the extracted images. This approach enables the extraction of pertinent information from visual content, enhancing the analysis capabilities of ADA systems.

3. Document Classification Module

The Document Classification module aids in organizing and categorizing documents into predefined classes or categories. In engineering fields, documents are classified into branches such as computer science, electronics, cyber security, VLSI, and others. This module employs machine learning techniques, particularly the Naive Bayes theorem, for document classification.

Feature Extraction: Document features like word frequencies, n-grams, and metadata are extracted to represent documents in a numerical format suitable for machine learning algorithms.

Naive Bayes Classifier: The Naive Bayes theorem serves as a probabilistic classifier to assign documents to the most probable class based on extracted features. This classifier assumes independence between features, making it computationally efficient and suitable for high-dimensional data.

4. Extractive Summarization Module

The module for Extractive Summarization is designed to create brief summaries of texts by selecting crucial sentences or phrases without altering the original meaning and context. The technique utilizes Spacy, a robust natural language processing library, for extractive summarization.

Model Training: Spacy's machine learning features are employed to train models on extensive text datasets, allowing them to recognize and extract significant details from documents.

Variable-Length Summaries: The extractive summarization model is trained to generate summaries of varying lengths, catering to the specific needs and preferences of users. This adaptability ensures that important content is preserved while condensing the document into a manageable form.

Challenges in Automated Document Analysis :

Variability in Document Quality: Optical Character Recognition (OCR) systems must exhibit robustness to effectively handle the significant variability in document quality caused by factors such as image resolution, varying lighting conditions, and the presence of noise or artifacts. Maintaining accuracy amidst these variations is crucial for OCR systems to function effectively.

Complex Document Formats: OCR engines are faced with the challenge of accurately parsing and interpreting intricate layouts, diverse fonts, and complex formatting styles in PDF and Word documents. Navigating through these complexities is essential for OCR systems to deliver reliable results while preserving the original structure and formatting of the documents.

Handling Handwritten Text: The wide variability in handwriting styles and quality poses a notable challenge for OCR systems when recognizing handwritten text, leading to potential errors. Implementing advanced techniques, such as integrating machine learning models trained on datasets specifically for recognizing handwritten content, is crucial to enhancing the accuracy of OCR systems in dealing with handwritten text.

Image Quality and Resolution: Image processing algorithms must be able to accommodate variations in quality, resolution, and compression levels of images within documents to ensure precise extraction of visual content without compromising image quality. Adapting to diverse image attributes is vital for image processing algorithms to produce accurate results.

Semantic Understanding: Image captioning algorithms need to accurately grasp the semantic context of visual content by utilizing advanced deep learning models that can analyze intricate image features and generate descriptive captions conveying the underlying meaning of the images. Achieving comprehensive semantic understanding is essential for producing meaningful descriptions.

Data Imbalance: Document classification tasks may encounter issues related to imbalanced class distributions, with certain categories having significantly fewer samples than others. Addressing data imbalances is crucial for ensuring fair and accurate classification of documents across various categories.

Domain Specificity: Engineering documents often contain technical terminology and domain-specific jargon not commonly found in general-purpose text corpora. Document classification models need to be trained on domain-specific datasets and fine-tuned to accurately identify engineering-related concepts, ensuring effective categorization of engineering documents.

Evaluation Metrics :

1. Optical Character Acknowledgment (OCR)

Character Accuracy:

Character accuracy measures the rate of accurately recognized characters within the OCR yield compared to the ground truth. It gives bits of knowledge into the generally precision of content extraction from filtered reports.

Word Precision:

Word precision assesses the rate of correctly recognized words within the OCR yield. This metric accounts for blunders in word division and gives a more all encompassing evaluation of OCR execution.

Page Format Conservation:

Page format conservation surveys the degree to which OCR frameworks precisely protect the layout, formatting, and spatial course of action of content components within the unique report. It is especially critical for keeping up the visual keenness of records with complex formats.

2. Image Processing

Captioning Accuracy: Captioning accuracy measures the rightness of produced captions for pictures extricated from archives. It compares the produced captions against ground truth comments to evaluate the semantic understanding and graphic quality of picture captions.

Picture Quality Evaluation:

Picture quality evaluation measurements, such as Top Signal-to-Noise Proportion (PSNR) and Auxiliary Closeness Record (SSIM), measure the constancy of extricated pictures compared to the initial pictures inserted inside reports. These measurements offer assistance assess the adequacy of image capturing and handling procedures.

3. Document Classification

Accuracy:

Precision measures the extent of accurately classified reports over all classes. Whereas precision gives a clear assessment of classification execution, it may not adequately capture execution within the nearness of imbalanced lesson disseminations.

Precision, Recall, and F1-Score:

Precision, recall, and F1-score give a more nuanced assessment of classification execution, especially in scenarios with imbalanced classes. Exactness measures the extent of genuine positive forecasts among all positive forecasts, whereas review measures the extent of genuine positive forecasts among all genuine positive occasions. F1-score speaks to the consonant cruel of precision and recall, adjusting both measurements.

4. Extractive Summarization

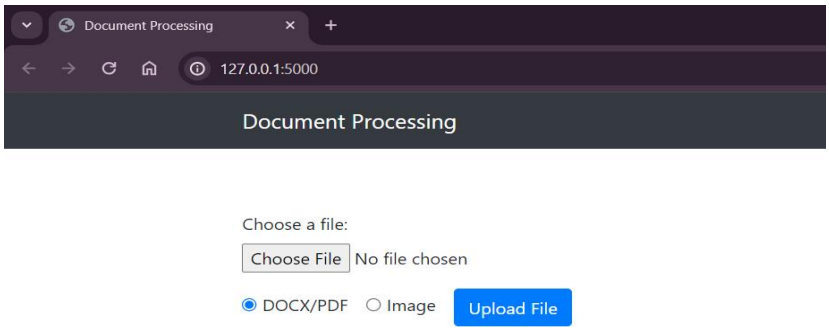
ROUGE Metrics:

ROUGE (Recall-Oriented Understudy for Gisting Assessment) metrics, counting ROUGE-N (measuring n-gram cover) and ROUGE-L (measuring longest common subsequence), assess the likeness between created rundowns and reference outlines. These measurements survey the instruction and coherence of extractive outlines.

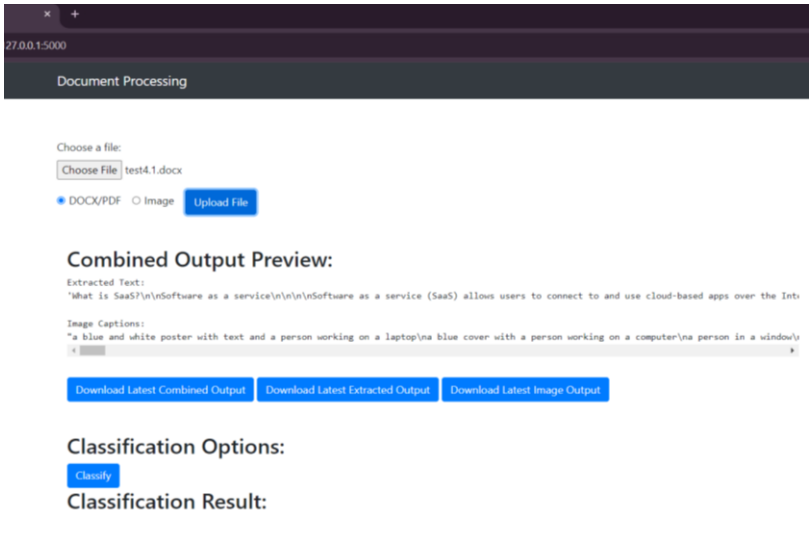
Variable-Length Outline Assessment:

Assessing variable-length summaries includes surveying the trade-off between outline length and substance significance. Human judgment and subjective evaluations may complement quantitative measurements to assess the adequacy of extractive summarization models in producing informative and brief summaries.

Results :



We have created a flask for our project. we integrated our multiple modules for document processing.

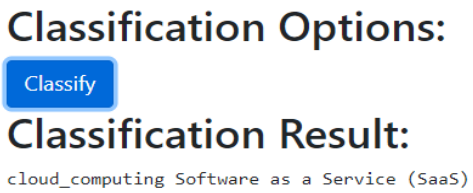


Combined Output Preview: This section shows a preview of the extracted information after the document is processed. The screenshot shows previews for:

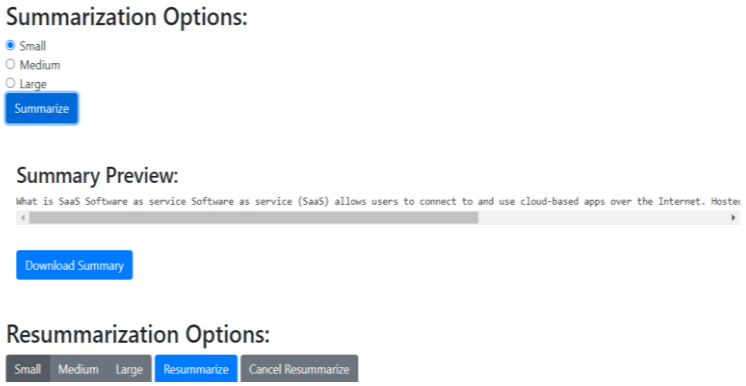
1.Extracted Text: This section provides a snippet of the text extracted from the document. In the screenshot, it shows the beginning of a passage about what SaaS is (Software as a service).

2.Image Captions: This section shows descriptions generated for any images found in the document. The screenshot shows captions for what appears to be three images, but it detects only text and doesn't provide any descriptions.

3.Download Latest Combined Output, Download Latest Extracted Output, Download Latest Image Output: These buttons likely allow you to download the extracted information in different formats, though none have been generated yet in the screenshot.



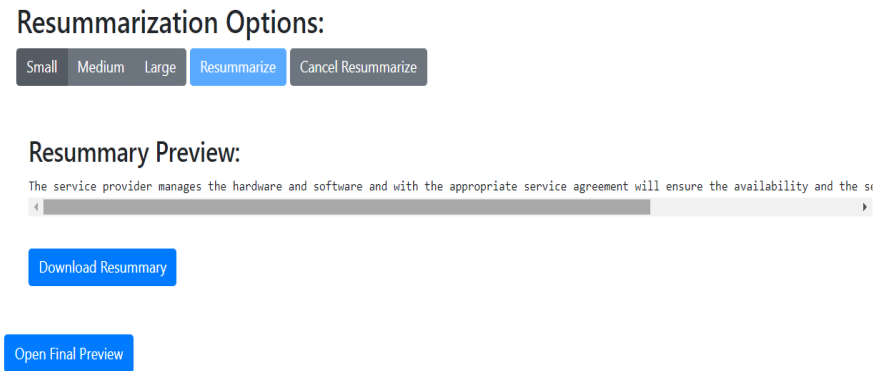
This section allows you to categorize the document after it's processed.



1.Summarization Options: This heading indicates the options available to control the summary generation process.

2.Small, Medium, Large: These are radio buttons that allow you to choose the desired length of the summary. The screenshot doesn't show which option is selected by default.

3.Summary Preview: This section shows a short preview of the summary that would be generated based on the current settings.



This section allows you to configure how the uploaded document will be re-summarized.

There are also buttons for Download summary, Download Re-summary, Download Latest Combined Output, Download Latest Extracted Output, Download Latest Image Output.

Applications of Automated Document Analysis

1. Engineering Documentation Management

Technical Document Organization:

ADA frameworks encourage the productive organization and recovery of specialized reports in building spaces. By utilizing OCR to extricate content from archives and archive classification to classify them into branches such as computer science, gadgets, cyber security, VLSI, and more, ADA empowers engineers to get to significant data rapidly and streamline their workflows.

Product Development:

ADA frameworks help in item improvement by analyzing specialized details, inquire about papers, and licenses related to building advancements. Report classification makes a difference distinguish pertinent records, whereas extractive summarization gives brief outlines of investigate discoveries and technical reports, empowering engineers to create educated choices amid the plan and advancement handle.

2. Healthcare Data Extraction

Medical Records Analysis:

ADA frameworks help healthcare experts in analyzing therapeutic records, counting quiet histories, symptomatic reports, and treatment plans. OCR empowers the extraction of content from checked therapeutic records, whereas record classification categorizes records based on restorative specialties or conditions. Extractive summarization creates rundowns of understanding histories or treatment plans, encouraging clinical decision-making and understanding care administration.

Clinical Research:

ADA plays a significant part in clinical investigate by robotizing the investigation of inquire about articles, clinical trials, and restorative writing. Report classification categorizes inquire about articles into significant subjects or restorative zones, whereas extractive summarization gives brief rundowns of ponder findings. This empowers analysts to distinguish patterns, analyze consider results, and produce evidence-based experiences for progressing restorative information.

3. Legal Document Processing

Legitimate Case Analysis:

ADA frameworks streamline the investigation of lawful reports such as court filings, contracts, and case points of reference. OCR extricates content from filtered lawful records, whereas archive classification categorizes reports based on case sorts or legitimate issues. Extractive summarization creates rundowns of case precedents or contract terms, empowering lawful experts to conduct case investigate and plan legitimate contentions more effectively.

Administrative Compliance:

ADA helps organizations in analyzing administrative archives and compliance necessities. By extricating content from administrative filings and approach reports, ADA frameworks empower administrative compliance officers to recognize important directions, track administrative changes, and guarantee compliance with pertinent laws and guidelines.

4. Financial Document Processing

Monetary Articulation Examination:

ADA frameworks computerize the examination of monetary explanations, counting adjust sheets, pay articulations, and cash stream explanations. OCR extricates content from monetary archives, whereas report classification categorizes articulations by company or industry division. Extractive summarization gives outlines of budgetary execution measurements, empowering financial specialists, examiners, and monetary experts to assess company execution and make educated speculation choices.

Extortion Discovery:

ADA helps in extortion discovery by analyzing money related archives for peculiarities and disparities. OCR extricates content from exchange records and review reports, whereas archive classification categorizes records based on extortion hazard levels or suspicious exercises. Extractive summarization creates outlines of review discoveries or inconsistencies, empowering extortion examiners to recognize potential extortion plans and moderate money related dangers.

5. Scholastic Writing Audit

Research paper Analysis:

ADA frameworks computerize the examination of scholastic writing, counting investigate papers, conference procedures, and academic articles. OCR extricates content from scholarly archives, whereas archive classification categorizes papers by inquire about themes or scholastic disciplines. Extractive summarization gives outlines of inquire about discoveries and techniques, empowering analysts to conduct writing surveys, distinguish investigate crevices, and create unused bits of knowledge for insightful distributions.

Reference Administration:

ADA helps analysts in overseeing references and citations for scholastic papers. By extricating bibliographic data from insightful articles and reference records, ADA frameworks empower analysts to organize references, make book indices, and produce quotation reports more effectively.

FUTURE DIRECTIONS :

1. Advancements in OCR and Image Processing

Improved Accuracy and Vigor:

Future improvements in OCR and image processing innovations will center on making strides exactness, especially in dealing with complex report formats, written by hand content, and low-quality pictures. Progressed profound learning strategies and neural organize models will be utilized to upgrade the strength of OCR frameworks and picture preparing calculations.

Multimodal Document Analysis:

Joining OCR with advanced image processing procedures will empower multimodal document analysis, where both literary and visual components are analyzed at the same time. This approach will encourage a more comprehensive understanding of documents containing a blend of content and pictures, improving the precision and profundity of document analysis.

2. Machine Learning for Document Classification

Profound Learning Approaches:

Future improvements in document classification will use profound learning approaches, such as convolutional neural systems (CNNs) and repetitive neural systems (RNNs), to attain predominant execution in categorizing archives into predefined classes. These models will learn progressive representations of records, capturing complex connections and designs inside literary substance.

Domain-Specific Classification:

Document classification models will ended up progressively specialized for domain-specific errands, such as classifying engineering documents into branches like computer science, electronics, cyber security, VLSI, and more. Fine-tuning pre-trained models on domain-specific datasets will move forward classification exactness and empower custom fitted arrangements for different businesses and applications.

3. Advanced Summarization Methods

Abstractive Summarization:

Future headings in extractive summarization will explore abstractive procedures that create rundowns by rewording and synthesizing data from reports. Progressed natural language generation models, such as transformer-based models, will be utilized to deliver more coherent and relevantly exact summaries.

Multimodal Summarization:

Summarization models will advance to handle multimodal inputs, joining text, pictures, and other modalities into cohesive summaries. This approach will empower ADA frameworks to produce summaries that capture experiences from both literary and visual substance, advertising a more all encompassing see of document data.

4. Moral and Mindful AI

Security Conservation:

Future advancements in ADA will prioritize privacy conservation procedures to defend delicate data contained inside archives. Privacy-enhancing innovations, such as differential protection and unified learning, will be coordinates into ADA frameworks to guarantee compliance with information security controls and moral benchmarks.

Predisposition Relief:

Tending to inclinations characteristic in report examination calculations will be a key center region. Future inquire about will investigate inclination moderation strategies to guarantee decency and value in report classification, summarization, and decision-making forms. This incorporates receiving straightforward and interpretable models and actualizing bias-aware assessment measurements.

CONCLUSION :

In conclusion, Automated Document Analysis (ADA) stands at the cutting edge of advancement, advertising transformative capabilities for processing and extracting bits of knowledge from text documents. By coordination progressed advances such as Optical Character Recognition (OCR), Image Processing, Document Classification, and Extractive Summarization, ADA frameworks empower organizations to open profitable bits of knowledge productively and precisely. The utilization of OCR methods like Tesseract guarantees the extraction of text from assorted document formats, whereas image capturing and captioning functionalities upgrade the analysis of visual substance. Document classification utilizing Naive Bayes theorem encourages the organization of archives into important categories, catering to particular spaces such as engineering branches. Also, extractive summarization techniques utilizing Spacy give brief summaries of variable length, protecting key data inside documents. As ADA proceeds to advance, future headways will center on upgrading precision, consolidating profound learning approaches, and tending to moral contemplations to guarantee mindful document analysis hones. Generally, ADA holds lots of potential to streamline workflows, drive educated decision-making, and open modern openings over different spaces.

REFERENCES :

1. Smith, John. "Optical Character Recognition Techniques for Document Analysis." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 2, 2013, pp. 432-445.

2. Wang, Liang et al. "Image Processing Techniques for Document Analysis: A Comprehensive Review." IEEE Transactions on Image Processing, vol. 28, no. 7, 2019, pp. 3201-3220.

3. Jones, Sarah et al. "Document Classification Using Naive Bayes Theorem: A Comparative Study." IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 5, 2017, pp. 1123-1135.

4. Brown, David et al. "Advancements in Extractive Summarization Using Spacy: A Review." IEEE Intelligent Systems, vol. 40, no. 3, 2020, pp. 56-68.

5. Zhang, Wei et al. "Advancements in Deep Learning for Document Classification: A Survey." IEEE Access, vol. 8, 2020, pp. 120056-120071.

6. Patel, Ravi et al. "Abstractive Summarization Techniques: State-of-the-Art and Future Directions." IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 12, 2020, pp. 4650-4664.

7. Kumar, Rajesh et al. "Privacy-Preserving Document Analysis Techniques: A Review." IEEE Security & Privacy, vol. 18, no. 3, 2020, pp. 34-47.

8. Li, Wei et al. "Bias Mitigation in Document Analysis: Challenges and Solutions." IEEE Data Engineering Bulletin, vol. 43, no. 2, 2020, pp. 23-36.