# PDF Extraction API Evaluation Template

**Team: TEAM-1**

**Team members:Sai Surya Madhav Rebbapragada** ,**Uday Kiran Dasari, Venkat Akash Varun Pemmaraju**

**Link to your analysis: https://github.com/DAMG7245-Big-Data-Sys-SEC-02-Fall24/Assignment2\_team1**

**Summary:**

PDF.co is best used for automating PDF processing tasks such as data extraction, conversion, editing, and form filling. It's particularly useful for businesses dealing with large volumes of PDF documents, needing to integrate PDF processing into their workflows, or requiring OCR capabilities. PDF.co excels in scenarios where custom PDF automation is needed, offering API integration and no-code options through platforms like Zapier and Make. However, it may not be the ideal choice for simple, one-off PDF tasks that don't require automation or for users who prefer desktop applications over cloud-based solutions.

## 1. General Information

| Attribute | Details |
| --- | --- |
| API Name | PDF.CO |
| Vendor | PDF.CO |
| Version/Release Date | February 2023 |
| Pricing Model | Free trail , Subscription, and Custom |
| Licensing and Compliance | HIPAA compliance support is available,GDPR compliance is in progress |

## 2. Technical Capabilities

| Feature | PDF.CO | Microsoft Document Intelligence | Other Vendor | Other Vendor |
| --- | --- | --- | --- | --- |
| File Format Support (PDF, DOCX, etc.) | Supports PDF to Excel, CSV, XML, JSON, Text, HTML, and Image conversion |  |  |  |
| OCR (Optical Character Recognition) | Accuracy of OCR for extracting text from images is below average for free trial api |  |  |  |
| Table Extraction | Ability to accurately detect and extract tables is good but could be better |  |  |  |
| Form Extraction | Handling of structured forms is not good, unable to handle it |  |  |  |
| Complex Layout Support | Ability to extract data from complex layouts (columns, images, embedded objects) is fine, but struggle with images for free trial api |  |  |  |
| Multi-language Support | Support for extracting content in multiple languages is not supported |  |  |  |
| Scalability and Performance | Speed and ability to scale across large datasets is very good |  |  |  |
| API Integration and Usability | Ease of API integration is difficult, there is no sdk provided , only endpoints are provided |  |  |  |
| Customization Options | Not available for Free trial Plan |  |  |  |
| Accuracy and Error Handling | It’s accurate around 60% but in some cases it extracts in even normal text to occasionally |  |  |  |

## 3. Business and Strategic Considerations

| Evaluation Metric | Adobe PDF API | Microsoft Document Intelligence | Other Vendor | Other Vendor |
| --- | --- | --- | --- | --- |
| Cost Efficiency (Pricing vs. Features) | Balance between cost and the number of features is decent compared to other providers |  |  |  |
| Vendor Reputation and Stability | a strong reputation in the market and demonstrated stability as a vendor, used by CNET & Wall Street Journal |  |  |  |
| Customer Support and SLA | Quality and response times of customer service is not extensively tested but many customers appreciated it. |  |  |  |
| Security and Privacy | PDF.co employs robust security measures including SSL/TLS encryption, user-controlled AES encryption, and secure Amazon AWS infrastructure. The platform supports HIPAA compliance, is working towards GDPR compliance, and offers features like data anonymization, IP restrictions, and secure file handling to address regulatory requirements such as GDPR and CCPA. |  |  |  |
| Documentation and Training Resources | Availability of user guides, developer resources is cluttered could be organized in a better way |  |  |  |
| Community and Ecosystem | PDF.co has a large user community with thousands of companies utilizing its services globally. The platform offers extensive third-party integrations, including popular tools like Zapier and Make, allowing connections with over 6,000 apps, and provides comprehensive developer resources to support its ecosystem |  |  |  |
| Roadmap and Innovation | The company continually enhances its platform with AI-powered capabilities, such as invoice parsing and document classification, indicating an ongoing focus on advanced API offerings. |  |  |  |
| Vendor Lock-in Risk | Not extensively test, but since it uses RestAPI based api, Migration is easy |  |  |  |

## 4. Performance Metrics

| Metric | Adobe PDF API | Microsoft Document Intelligence | Other Vendor | Other Vendor |
| --- | --- | --- | --- | --- |
| Latency | While it depends with respect to size of pdf, usually it is around 20-30 seconds |  |  |  |
| Throughput | Has Not been extensively tested yet. |  |  |  |
| Error Rate | Encountered various error types including HTTP status codes (400, 401, 402, 403, 408), file access issues, URL-related errors, SSL certificate problems, and timeout errors for large documents. While specific error frequencies aren't provided, the platform offers detailed troubleshooting guides and solutions to mitigate these common issues, such as using asynchronous mode and adjusting access parameters. |  |  |  |
| Data Loss/Integrity | rough estimate for the percentage of lost or misinterpreted data during extraction using PDF.co's API could be between 5% to 15% |  |  |  |

## 5. Value-Add Features

| Feature | Adobe PDF API | Microsoft Document Intelligence | Other Vendor | Other Vendor |
| --- | --- | --- | --- | --- |
| Advanced AI/ML Capabilities | AI/ML support for better contextual extraction, document understanding. |  |  |  |
| Pre-built Templates for Specific Use Cases | Availability of industry-specific extraction templates (e.g., legal, financial). |  |  |  |
| Document Classification/Tagging | Auto-classification of documents based on content. |  |  |  |
| Metadata Extraction | Ability to extract embedded metadata (author, timestamp, etc.). |  |  |  |

## 6. Overall Evaluation

| Attribute | Rating | Comments |
| --- | --- | --- |
| Technical Fit | Rate out of 10 | Suitability of API based on technical capabilities. |
| Business Fit | Rate out of 10 | Suitability based on business requirements. |
| Total Cost of Ownership | Rate out of 10 | Long-term cost, including licensing, support, and infrastructure. |
| Ease of Implementation and Use | Rate out of 10 | Developer experience, integration challenges. |
| Vendor Reliability and Support | Rate out of 10 | Overall impression of vendor stability and support. |

## 7. Recommendations

| Recommendation | Details |
| --- | --- |
| Best Fit for the Use Case | Summary of which API(s) are most suitable for the given use case. |
| Further Considerations | Any additional notes, alternative solutions, or potential challenges. |