*IFC_2.0_Q1_*Advance-Text-Analytics-for-Entity-and-Classification

Theme	Intelligent Data Processing & Enrichment
Initiative	Advance Text Analytics for Entity & Classification
Document ID	PRD-IFC-T2.2-V1.0
Owner	Anuj S.
Team	Data Science Team, Engineering Team, QA Team
Summary	Enhance the system's ability to automatically extract key entities and classify documents/text segments, improving data enrichment for search and analysis.
Status	Discovery / Delivery
Next Milestone	Text Analytics Model Training Completion (Q3 2026)

Requirement Definition

Objective

To significantly improve the automatic identification and categorization of key information (entities like persons, locations, units, events, dates) and document-level classifications from all text-based data. This will make data more structured, discoverable, and valuable for Data Analysts, aligning with "Actionable Intelligence Generation."

Background context

While basic text extraction is in place, the system currently lacks sophisticated automated text analytics capabilities. Analysts spend considerable time manually identifying key entities and categorizing documents. Implementing advanced text analytics will automate this laborious process, providing richer metadata that can be leveraged for advanced search, visualization, and predictive modeling.

Problems

- **Manual Data Tagging:** Analysts manually identify and tag entities and classifications, which is time-consuming and inconsistent.
- **Limited Search Granularity:** Search is less effective without structured entity and

classification metadata.

- **Underutilized Unstructured Data:** Valuable insights are buried in large volumes of unstructured text.

Constraints

- **Training Data:** Availability and quality of labeled datasets for training domain-specific NLP models.
- **Computational Resources:** Text analytics can be computationally intensive, requiring sufficient processing power.
- **Model Maintenance:** Ongoing effort required to maintain and update NLP models as data evolves.
- **Accuracy vs. Performance:** Balancing high accuracy requirements with acceptable processing times.

Solution briefing

We will develop and integrate advanced Natural Language Processing (NLP) models for entity extraction and text classification. These models will be trained on domain-specific data to ensure high relevance and accuracy. The output of these models will be stored as structured metadata in the Big Data Repository and integrated with the search index.

Functional requirements (User Stories with Acceptance Criteria)

- **User Story 4.2.1:** As a System, when text-based data is ingested or OCR'd, I want to automatically trigger the Text Analytics module as part of the workflow, so that data can be enriched. (Ref. FR-TA-001.1, FR-WFA-001)
  - **Acceptance Criteria:**
    - **AC 4.2.1.1:** GIVEN a text-based document (e.g., digital document, OCR-processed scanned document) is successfully stored in the BDR, WHEN the ingestion workflow reaches the Text Analytics stage, THEN the Text Analytics module SHALL be automatically invoked.
    - **AC 4.2.1.2:** GIVEN the Text Analytics module is invoked, THEN it SHALL process the document's text content.
- **User Story 4.2.2:** As a System, I want to automatically classify documents or specific text segments into predefined categories (e.g., "Financial Report," "Operational Briefing," "Threat Assessment"), so that documents can be easily categorized and filtered. (Ref. FR-TA-001.2)
  - **Acceptance Criteria:**
    - **AC 4.2.2.1:** GIVEN a document's text content, WHEN the Text Analytics module processes it, THEN it SHALL assign one or more predefined classification tags (e.g., "Operational," "Logistics," "Personnel") to the document.
    - **AC 4.2.2.2:** GIVEN a document is classified, THEN its classification tags SHALL be stored as metadata in the BDR and indexed for search.
    - **AC 4.2.2.3:** GIVEN a document's content, WHEN it clearly belongs to a specific category (e.g., "Financial Report"), THEN the system SHALL correctly classify it

with 90% accuracy.

- **User Story 4.2.3:** As a System, I want to automatically identify and extract specific types of entities (e.g., persons, organizations, units, locations, events, dates) from text, so that key information is structured and searchable. (Ref. FR-TA-001.4)
    - **Acceptance Criteria:**
        - **AC 4.2.3.1:** GIVEN a document's text content, WHEN the Text Analytics module processes it, THEN it SHALL identify and extract entities of types: "Person," "Organization," "Unit," "Location," "Event," and "Date."
        - **AC 4.2.3.2:** GIVEN an extracted entity, THEN it SHALL be categorized by its type and stored as structured metadata in the BDR, linked to the original document and its position in the text.
        - **AC 4.2.3.3:** GIVEN a document containing clear instances of these entity types, WHEN processed, THEN the entity extraction shall achieve 85% precision and recall for each specified type.
- **User Story 4.2.4:** As a Data Analyst, I want to view the extracted classifications and entities associated with a document or record, so that I can quickly understand its key contents. (Implicit from FR-TA-001.5, UI aspect)
    - **Acceptance Criteria:**
        - **AC 4.2.4.1:** GIVEN a Data Analyst views a document in the system's document viewer, THEN a dedicated panel (e.g., RHS panel) SHALL display a list of all extracted classifications and entities for that document.
        - **AC 4.2.4.2:** GIVEN the extracted entities are displayed, WHEN the Data Analyst clicks on an entity in the panel, THEN the corresponding text in the document viewer SHALL be highlighted.
- **User Story 4.2.5:** As a Data Analyst, I want to be able to search and filter data using the extracted classifications and entities, so that I can perform more precise analytical queries. (Ref. FR-TA-001.5, FR-SRCH-001)
    - **Acceptance Criteria:**
        - **AC 4.2.5.1:** GIVEN extracted classifications are stored, WHEN a Data Analyst uses the search interface, THEN the search interface SHALL provide filters for document classifications (e.g., a dropdown of categories).
        - **AC 4.2.5.2:** GIVEN extracted entities are stored, WHEN a Data Analyst uses the search interface, THEN the search interface SHALL provide filters for specific entity types and values (e.g., "Location: New Delhi," "Person: John Doe").
        - **AC 4.2.5.3:** GIVEN a search is performed using classification/entity filters, THEN the results SHALL accurately reflect documents matching those specific metadata attributes.

Non-functional requirements

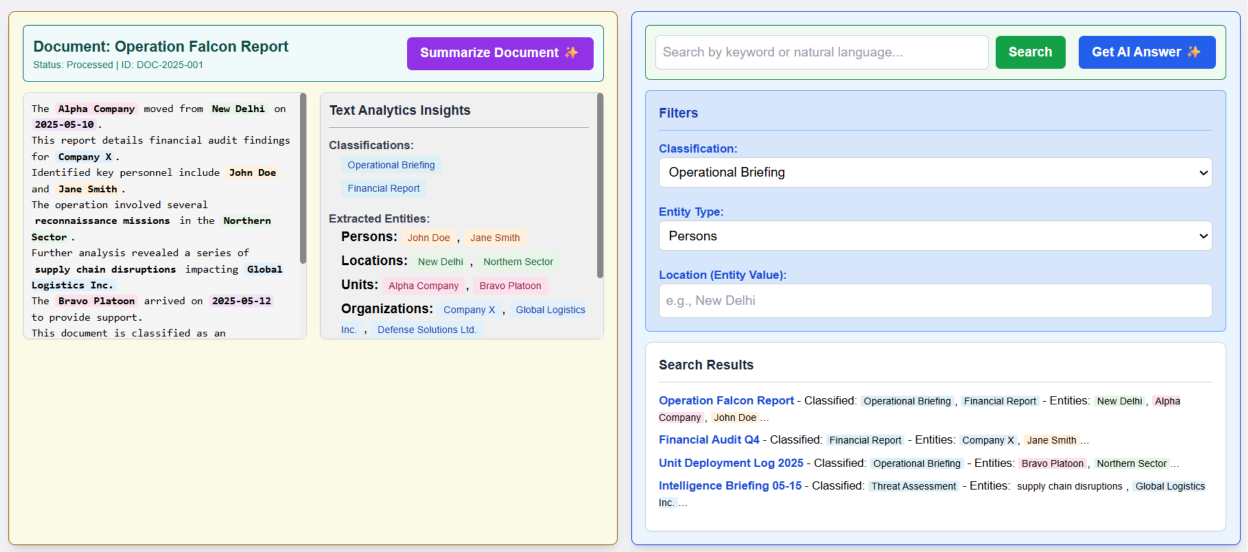
- **NFR-TA-ACC-001:** Text classification shall achieve a minimum of **90% accuracy** for predefined categories.
- **NFR-TA-ACC-002:** Entity extraction shall achieve a minimum of **85% precision and recall** for core entity types (persons, locations, units, organizations).

- **NFR-TA-PER-001:** Text analytics processing (classification and entity extraction) for a standard document (e.g., 10 pages) shall complete within 15 seconds.
- **NFR-TA-SCL-001:** The Text Analytics module shall be scalable to process concurrently ingested documents without becoming a bottleneck, handling up to 50 documents/minute.
- **NFR-TA-REL-001:** The Text Analytics module shall gracefully handle malformed or empty text inputs without crashing, logging errors and proceeding.

Designs

Wireframes, mockups and data flow diagrams as they are developed, showing what the feature will look like, and how users will interact with it.

Mockup link: https://g.co/gemini/share/fbf0e3bc26a8



Document: Operation Falcon Report

Status: Processed | ID: DOC-2025-001

```
The Alpha Company moved from New Delhi on
2025-05-10 .
This report details financial audit findings
for Company X .
Identified key personnel include John Doe
and Jane Smith .
The operation involved several
 reconnaissance missions in the Northern
Sector .
Further analysis revealed a series of
 supply chain disruptions impacting Global
Logistics Inc.
The Bravo Platoon arrived on 2025-05-12
to provide support.
This document is classified as an
```

Text Analytics Insights

Classifications:

Operational Briefing

Financial Report

Extracted Entities:

Persons: John Doe , Jane Smith

Locations: New Delhi , Northern Sector

Units: Alpha Company , Bravo Platoon

Organizations: Company X , Global Logistics Inc. , Defense Solutions Ltd.

Filters

Classification: Operational Briefing

Entity Type: Persons

Location (Entity Value):

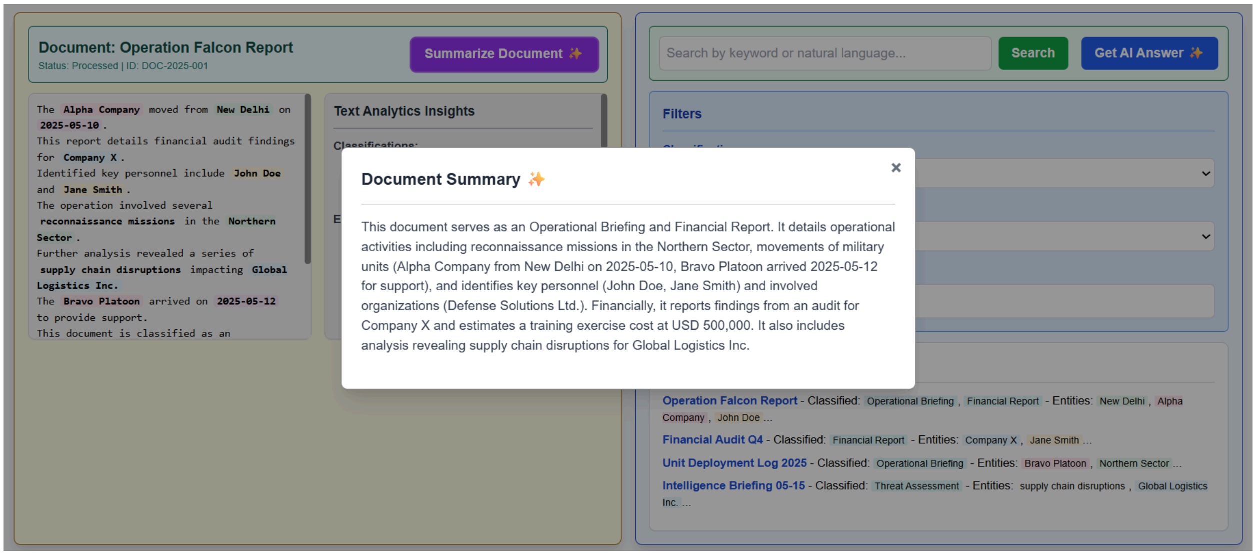
Search Results

Operation Falcon Report - Classified: Operational Briefing , Financial Report - Entities: New Delhi , Alpha Company , John Doe ...

Financial Audit Q4 - Classified: Financial Report - Entities: Company X , Jane Smith ...

Unit Deployment Log 2025 - Classified: Operational Briefing - Entities: Bravo Platoon , Northern Sector ...

Intelligence Briefing 05-15 - Classified: Threat Assessment - Entities: supply chain disruptions , Global Logistics Inc. ...



Document: Operation Falcon Report
Status: Processed | ID: DOC-2025-001

The Alpha Company moved from New Delhi on 2025-05-10 .
This report details financial audit findings for Company X .
Identified key personnel include John Doe and Jane Smith .
The operation involved several reconnaissance missions in the Northern Sector .
Further analysis revealed a series of supply chain disruptions impacting Global Logistics Inc.
The Bravo Platoon arrived on 2025-05-12 to provide support.
This document is classified as an

Document Summary ✨

This document serves as an Operational Briefing and Financial Report. It details operational activities including reconnaissance missions in the Northern Sector, movements of military units (Alpha Company from New Delhi on 2025-05-10, Bravo Platoon arrived 2025-05-12 for support), and identifies key personnel (John Doe, Jane Smith) and involved organizations (Defense Solutions Ltd.). Financially, it reports findings from an audit for Company X and estimates a training exercise cost at USD 500,000. It also includes analysis revealing supply chain disruptions for Global Logistics Inc.

Operation Falcon Report - Classified: Operational Briefing , Financial Report - Entities: New Delhi , Alpha Company , John Doe ...

Financial Audit Q4 - Classified: Financial Report - Entities: Company X , Jane Smith ...

Unit Deployment Log 2025 - Classified: Operational Briefing - Entities: Bravo Platoon , Northern Sector ...

Intelligence Briefing 05-15 - Classified: Threat Assessment - Entities: supply chain disruptions , Global Logistics Inc. ...

Reference

- SRS-IFC-V1.0, Section 3.1.6 Text Analytics (FR-TA-001)
- IDD-IFC-V1.0, Screen: Document/Record View (with TA Insights)
- IDD-IFC-V1.0, Screen: Search Dashboard (Report-Specific Filters)

OKR metric

- **Objective:** Enhance the system's ability to automatically extract and classify key information.
- **Key Results:**
  - Achieve 85% precision/recall for key entity extraction (persons, locations, units) by end of Q3 2026.
  - Achieve 90% accuracy for automated document classification across defined categories by end of Q3 2026.
  - Increase user satisfaction with search capabilities by 15% (partially attributed to better TA metadata) by end of Q3 2026.

Stakeholders

- **Responsible:** Data Science Team, Engineering Team (Text Analytics Service, Search Integration)
- **Accountable:** Product Manager
- **Consulted:** Data Analysts, QA Team, User Experience Team
- **Informed:** Project Management, System Administrators

Risks

- **R-TA-001: Training Data Scarcity:** Lack of sufficient, high-quality, labeled training data for domain-specific models.
  - **Mitigation:** Prioritize data labeling efforts. Explore active learning techniques. Leverage pre-trained models and fine-tuning.

- **R-TA-002: Model Drift:** Accuracy degradation over time as data patterns change.
  - **Mitigation:** Implement continuous model monitoring. Establish a retraining pipeline.
- **R-TA-003: Performance Impact:** Processing large documents or high volumes could cause latency.
  - **Mitigation:** Optimize model inference. Utilize GPU acceleration. Scale out Text Analytics service instances.

Decision log

- **Question:** Which NLP framework/library to use for initial development?
  - **Decision:** Start with a combination of spaCy for rule-based/statistical NER and Hugging Face Transformers for more advanced classification/fine-tuning, due to flexibility and community support.
  - **Decider:** Data Science Lead, Engineering Lead
  - **Date:** 2025-06-01

Launch Readiness

Testing plan

- **Unit Testing:** For individual NLP model components, entity extraction rules, and classification algorithms.
- **Integration Testing:** Verify seamless data flow from ingestion to Text Analytics, and from Text Analytics to BDR and Search Index.
- **Functional Testing:** Execute test cases covering all user stories and acceptance criteria, using diverse document types and content.
- **Accuracy Testing:** Develop automated evaluation pipelines to measure precision, recall, and F1-score for entity extraction and classification accuracy against ground truth datasets.
- **Performance Testing:** Measure processing time for various document sizes and concurrent loads.
- **User Acceptance Testing (UAT):** Involve Data Analysts to validate the utility of extracted entities/classifications in search and their overall relevance.

Tracking & analytics

- **Text Analytics Dashboard:** Track accuracy metrics (precision, recall, F1-score) for each entity type and classification category.
- **Processing Time:** Monitor average processing time per document.
- **Search Usage:** Track usage of classification and entity filters in the search module.

Marketing plan

- Internal communication highlighting "smarter search" and "deeper insights" due to automated tagging.
- Training sessions for Data Analysts on how to leverage new search filters and interpret

extracted entities.

Customer service plan

- Brief customer service on expected accuracy levels and common scenarios for reporting incorrect extractions/classifications.

Legal checks

- Ensure compliance with data privacy regulations regarding the extraction and storage of sensitive entities (e.g., PII).

FAQs

- "What kinds of things can the system automatically identify now?"
- "How accurate is the automatic tagging?"
- "Can I search by location or person name directly?"

Impact

Success metrics

- **KPI:** 85% precision/recall for key entity extraction (persons, locations, units).
- **KPI:** 90% accuracy for automated document classification across defined categories.
- **KPI:** 20% increase in analyst efficiency for initial document triage.

Next steps

- Explore relationship extraction.
- Investigate automated summarization beyond LLM (e.g., extractive summarization).
- Expand entity types based on user feedback.