Architecting a Multi-Modal Pipeline for Scientific Document Audit and Quality Assessment

December 23, 2025

Abstract

The rapid proliferation of scientific literature necessitates the development of automated tools for content verification, structural analysis, and linguistic auditing. This report introduces a sophisticated, eight-phase NLP pipeline that processes raw scientific documents (PDFs and images) into high-level quality reports. We integrate Neural Optical Character Recognition (OCR), Deep Sequential Section Classification, Transformer-based Grammar Error Correction (GEC), and Factual Consistency Verification. By leveraging models such as Nougat, T5, and CNN-BiLSTM hybrids, the system provides a comprehensive "Quality Score" that evaluates manuscripts across structural, linguistic, and factual dimensions.

1 Introduction and Summary

The manual review of scientific manuscripts is an arduous process prone to human bias and inconsistency. The objective of this project was to construct a "Document Auditor" capable of understanding the visual layout of a paper, extracting its semantic structure, and verifying the rigor of its claims.

The framework is built on the principle of *Modular Intelligence*. It begins by ingesting a massive corpus of peer reviews (ASAP-Review and PeerRead) to learn the patterns of high-quality scientific discourse. It then utilizes visual-transformer models to parse raw input, cleans that input using fine-tuned generative models, and finally subjects the text to a factual consistency check against the FEVER (Fact Extraction and VERification) dataset logic. The end result is not merely a transcription, but a critical evaluation of the document's scholarly merit.

2 Detailed Methodology by Phase

2.1 Phase 1: Multi-Source Data Ingestion

The pipeline begins with the ingestion of the *ASAP-Review* and *PeerRead* datasets. These datasets provide the "gold standard" for how scientific sections (Introduction, Methodology, results) are typically structured and reviewed.

- **Preprocessing Strategy:** We implemented a custom regex-based cleaning engine to remove non-ASCII characters while preserving scientific notation.
- **Tokenization:** Using NLTK, we filtered for tokens between 2 and 15 characters, ensuring that mathematical variables and extreme outliers (OCR noise) were excluded from the vocabulary.

2.2 Phase 2: Tiered Semantic Embeddings

To represent scientific text, we moved beyond simple frequency-based models. We implemented a tiered embedding strategy to capture both local syntax and global semantics:

- **FastText Configuration:** To handle the "out-of-vocabulary" (OOV) problem common in specialized scientific domains, we trained a FastText model.
  Training Arguments: `vector_size=100, window=5, min_count=5, sg=0 (CBOW), epochs=10`.
- **Sentence-BERT (SBERT):** We utilized the `all-MiniLM-L6-v2` model to generate 384-dimensional contextual embeddings for entire paragraphs, allowing the system to understand the "intent" of a section.

2.3 Phase 3: Deep Sequential Section Classification

A core challenge in scientific NLP is distinguishing between "Related Work" and "Introduction." We developed a Hybrid CNN-BiLSTM model to solve this. The CNN layers act as "feature extractors" for specific n-gram indicators (e.g., "we propose," "state-of-the-art"), while the BiLSTM layers model the sequential flow.

- **Architectural Parameters:**
  - `Embedding Layer:` Input dimension 20,000, Output dimension 128.
  - `Conv1D:` 64 filters, kernel size 5, ReLU activation.
  - `Bidirectional LSTM:` 64 units per direction, Dropout=0.5.

- **Training Hyperparameters:** `Optimizer: Adam`, `Loss: Sparse Categorical Crossentropy`, `Batch Size: 32`, `Epochs: 10`.
- **Class Balancing:** We applied a `TARGET_SIZE=3000` per label via oversampling to prevent the "Methodology" class (the most common) from biasing the model.

2.4 Phase 4: Visual-to-Semantic Parsing (Nougat)

Unlike standard PDF parsers (like PyPDF2) which fail on complex LaTeX formatting and tables, we integrated **Nougat** (facebook/nougat-base). This model treats the document as an image and decodes it into a structured Markdown format.

- **Implementation:** We utilized the `VisionEncoderDecoderModel` with a greedy decoding strategy for stability.
- **Output:** The model produced `.mmd` files which were then parsed into a structured dictionary of headings and body text.

2.5 Phase 5: Generative Grammar Error Correction (GEC)

Scientific documents often contain non-native linguistic patterns or OCR-induced hallucinations. We fine-tuned a **T5-base** model on the JFLEG dataset to perform "Grammatical Error Correction" while preserving scientific meaning.

- **Training Arguments:** `Learning_rate=5e-5, Weight_decay=0.01, Warmup_ratio=0.1, Epochs=10, per_device_train_batch_size=4`.
- **Supervision Strategy:** We used "Multi-Reference Supervision," where the model randomly selects one of several valid human corrections during training to increase robustness.

2.6 Phase 6: Factual Consistency Verification

To detect internal contradictions, we implemented a Fact-Checker using **T5-small** fine-tuned on the *FEVER* dataset.

- **Task Design:** The model takes an "evidence" segment (e.g., from the Methodology) and a "claim" (e.g., from the Conclusion) and labels the relationship.
- **Training Arguments:** `Num_epochs=1, Learning_rate=5e-5, Batch_size=8, Max_input_length`
- **Inference:** Prefix-based prompting: `"verify: [evidence] claim: [hypothesis]"`.

3 Experimental Results and Analysis

3.1 Classification Benchmarks

The CNN-BiLSTM Hybrid demonstrated a validation accuracy of **89%**. The model showed particularly high precision in identifying "Experiments" (0.91) due to the presence of metric-heavy tokens (accuracy, p-value, %). The "GRU" model showed slightly lower performance (85%) but was 30% faster during training, suggesting a viable trade-off for real-time applications.

3.2 Grammar Correction Efficacy

Before GEC, OCR-extracted text often suffered from pluralization errors (e.g., "results demonstrates"). Post-processing with our fine-tuned T5 model improved the Google-BLEU (GLEU) score by approximately **14%**, indicating a transition from raw transcription to professional-grade scientific writing.

3.3 Integrated Quality Scoring

Using the Phase 8 Scoring system, we evaluated the *"Attention is All You Need"* paper as a test case. The results were as follows:

- **Structure Score (9.5/10):** All essential sections (Intro, Method, Results) were detected.
- **Order Score (10/10):** The sequence followed the logical Introduction $\rightarrow$ Conclusion flow.
- **Consistency Score (8.2/10):** Claims regarding "Self-Attention" were supported by the Methodology text extracted by Nougat.

4 Conclusion

The framework developed in this project successfully navigates the complexities of academic document analysis. By bridging the gap between computer vision (OCR) and high-level linguistic auditing (Fact-Checking), we have created a tool that provides objective feedback on scientific rigor.

Key takeaways include:

1. **Hybrid Architectures Matter:** Standard LSTMs are insufficient for scientific text; convolutional layers are necessary to capture the specific nomenclature of academic writing.

2. **Generative Post-Processing:** GEC is not just about grammar; it is essential for cleaning the noise inherent in automated document parsing.

3. **Fact-Checking as a Metric:** Internal consistency is the ultimate measure of paper quality, and Transformer models are now capable of detecting these nuances at scale.

Future research will focus on **Cross-Document Verification**, where the fact-checker compares claims against a broader knowledge graph (like Semantic Scholar) to identify plagiarism or lack of novelty.