Biomedical Document Classification Pipeline

Model Comparison Analysis

SapBERT vs Dual-Model (SapBERT + PubMedBERT)

Prepared by: Nalini Panwar, Lead Data Engineer
Date: December 2025

Executive Summary

This report presents empirical findings comparing single-model (SapBERT) versus dual-model (SapBERT + PubMedBERT) approaches for biomedical document classification.

KEY FINDING: The dual-model approach shows **-1.3% accuracy difference** compared to SapBERT alone. SapBERT actually **OUTPERFORMS** the dual model. The added complexity provides no benefit.

RECOMMENDATION: Use SapBERT-only

Background

The classification pipeline was initially designed with a dual-model architecture based on the hypothesis that:

- **SapBERT:** Optimized for short biomedical terms (trained on UMLS concept names)
- **PubMedBERT:** Better for longer narrative text (trained on PubMed abstracts)
- **Dual (70/30 fusion):** Could capture benefits of both models

Methodology

Test Data

87 biomedical categories were tested covering: Demographics, Reproductive, Lifestyle, Measurements, Informed Consent, Vitals, Clinical Labs, Assessments, and Document Structure domains.

Test data was extracted from a curated dataset of clinical trial protocol documents.

Test Case Types

Type	Count	Description	Example
Short	4,034	Brief headings, 1-5 words	"Inclusion Criteria", "ECOG"

Long	1,983	Full narrative sentences	"Patients must have hemoglobin..."
Ambiguous	5,653	Multi-topic or unclear	"Study Population and Eligibility"

Results

Overall Accuracy

Model	Accuracy	Difference vs SapBERT
SapBERT Only	28.2%	—
PubMedBERT Only	12.2%	-16.0%
Dual (70/30)	26.9%	-1.3%

Accuracy by Case Type

Case Type	SapBERT	PubMedBERT	Dual
Short Headings	47.8%	23.1%	44.1%
Long Narrative	10.4%	2.1%	10.3%
Ambiguous	20.5%	8%	20.4%

Speed Comparison

Metric	SapBERT Only	Dual Model
Inference Time	3019 ms	5959 ms
Overhead	—	+97% slower
Models to Load	1	2
Memory Usage	~1.5 GB	~3.0 GB

Analysis

Key Observations

1. **SapBERT outperforms Dual model:** SapBERT alone (28.2%) beats the dual approach (26.9%) by 1.3 percentage points across all case types.

2. **Short headings show strongest performance:** SapBERT achieves 47.8% accuracy on short headings, confirming its strength with biomedical terms.

3. **PubMedBERT significantly underperforms:** At only 12.2% accuracy, PubMedBERT alone is 16 percentage points worse than SapBERT.

4. **Dual model adds overhead without benefit:** The 70/30 fusion actually degrades SapBERT's performance while doubling inference time.

Recommendation

USE SAPBERT-ONLY FOR CLASSIFICATION

The dual-model approach is **NOT JUSTIFIED** based on empirical results:

Factor	Impact
Accuracy	-1.3% (Dual is WORSE than SapBERT alone)
Speed	+97% slower inference
Memory	2x GPU/RAM usage
Complexity	Additional code paths and failure modes

Conclusion

Empirical testing on **11,670 classification samples** across **87 biomedical categories** demonstrates that the single-model SapBERT approach outperforms the dual-model architecture in accuracy, speed, and resource efficiency.

Final Decision: Implement SapBERT-only classification pipeline.

Technical Notes

This technical analysis represents work completed for production deployment in a clinical document intelligence system. The methodology and results are based on real-world requirements for processing biomedical research documentation at scale. The analysis has been sanitized for public sharing while maintaining architectural integrity.