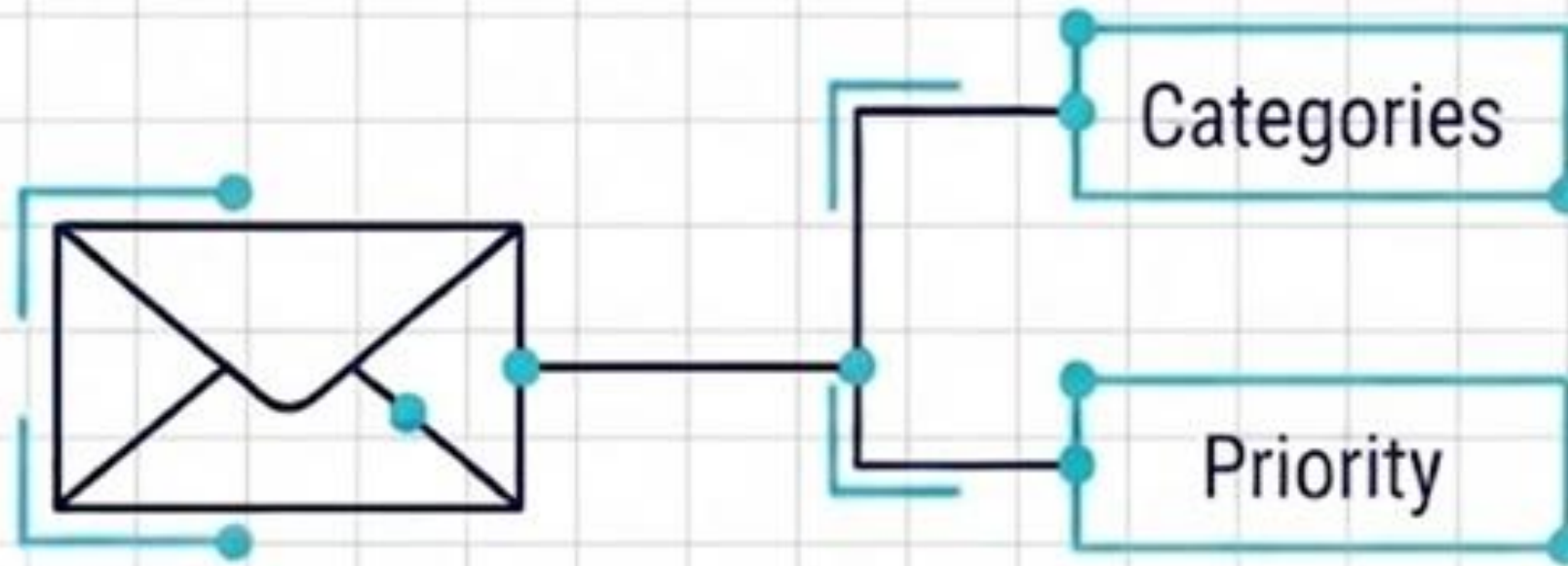


Enhancing Email Management through Machine Learning

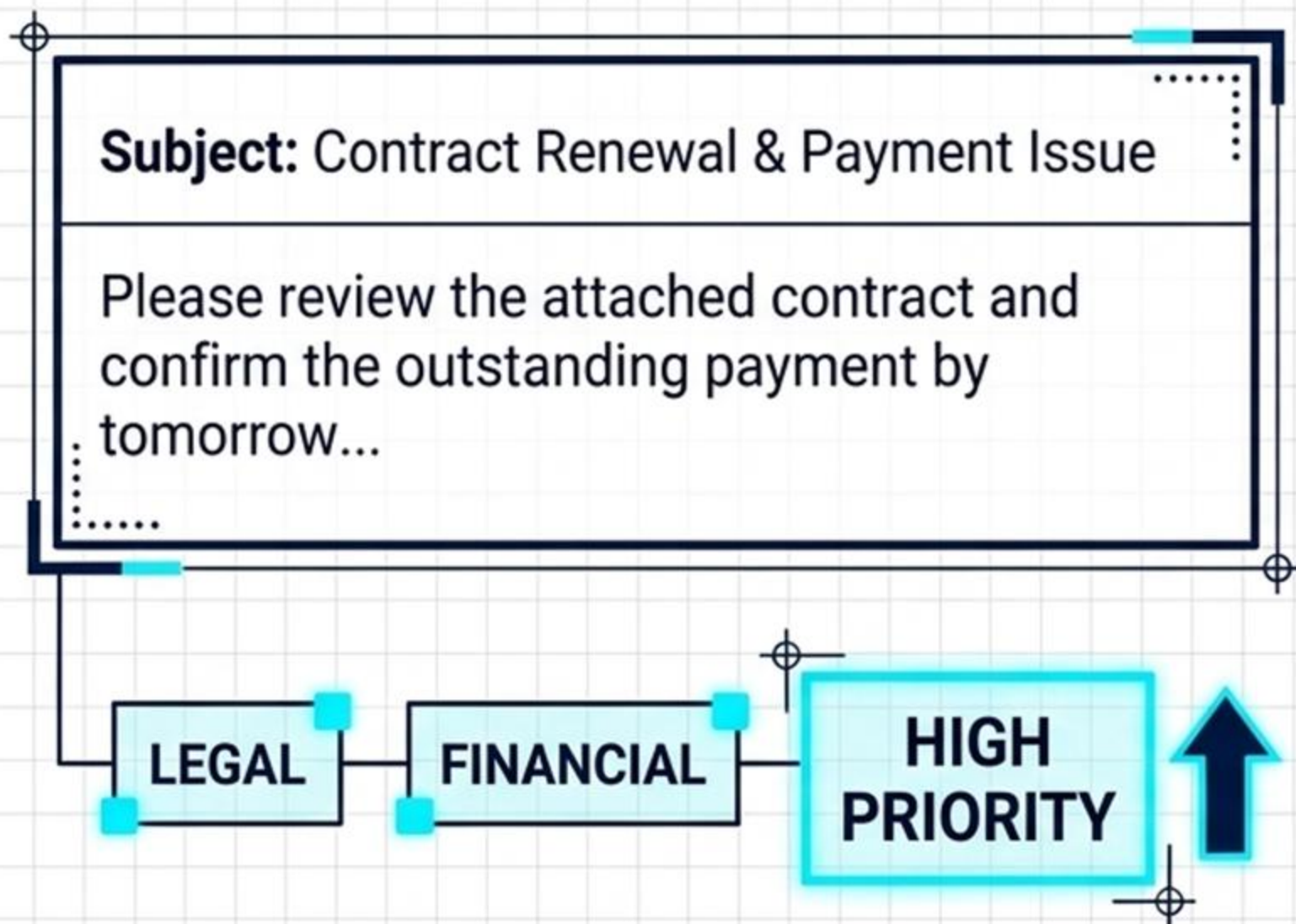
A Multi-Task DistilBERT Framework for Multi-Label Email Classification and Priority Prediction



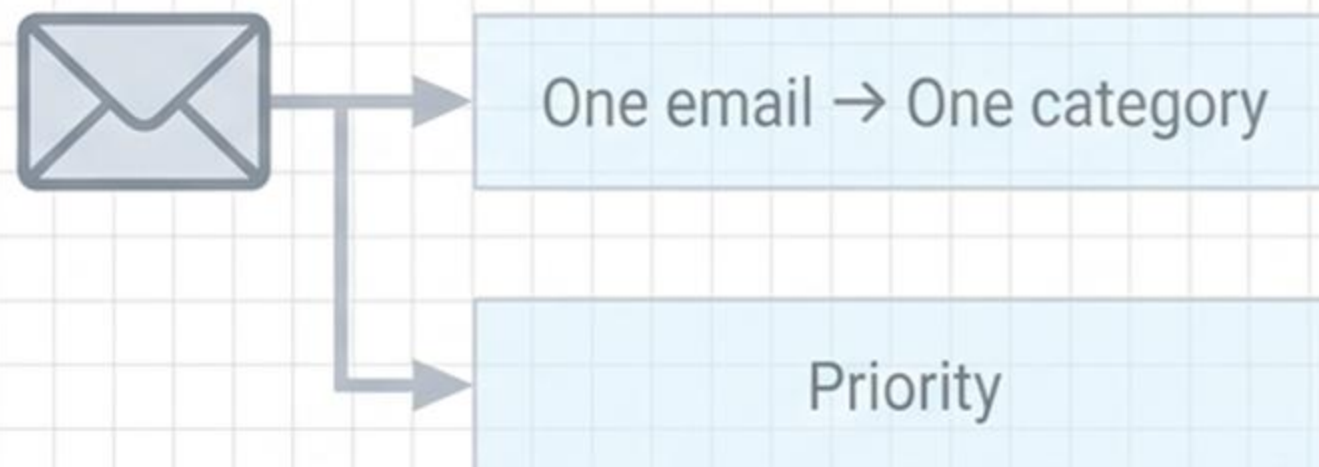
- **Researcher:** Abubaker Hussein
- **Degree:** MSc in Artificial Intelligence
- **Institution:** Midocean University

Thesis Advisor: Dr. Amjad Qatish
Defense Date: 15-8-2026

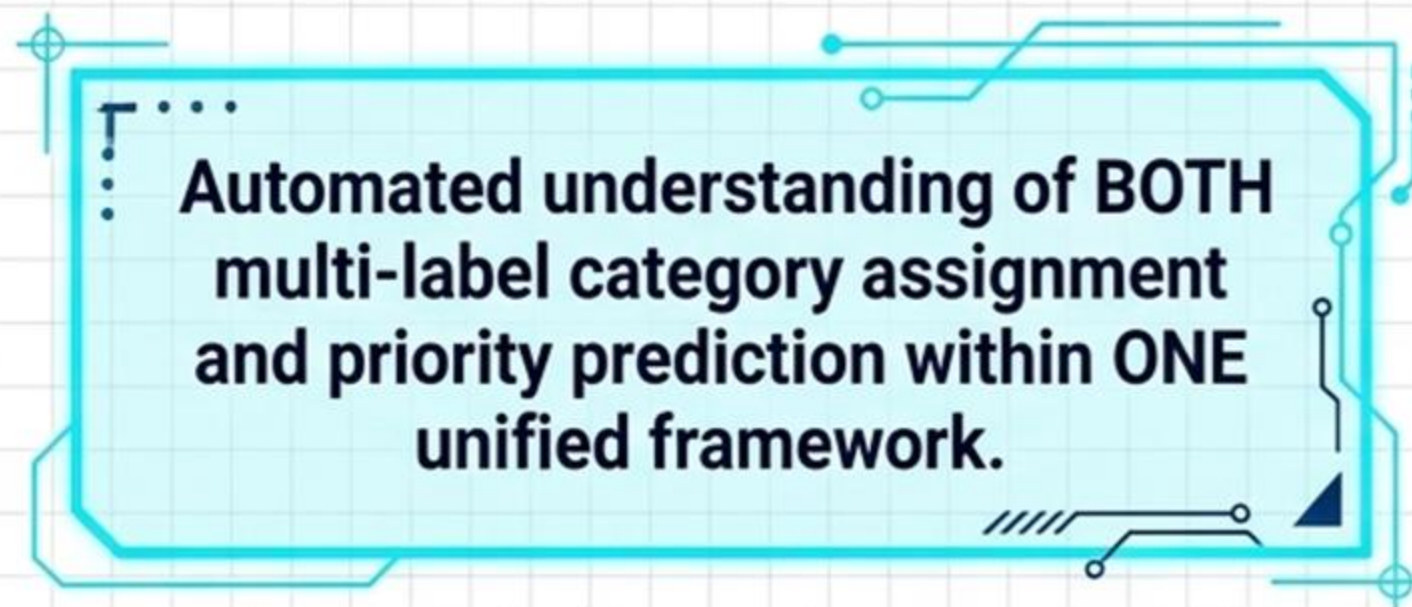
Organizational emails are heterogeneous and may contain multiple intents within one message. Manual organization is time-consuming and difficult to scale.



Traditional Flaw:



Research Need:



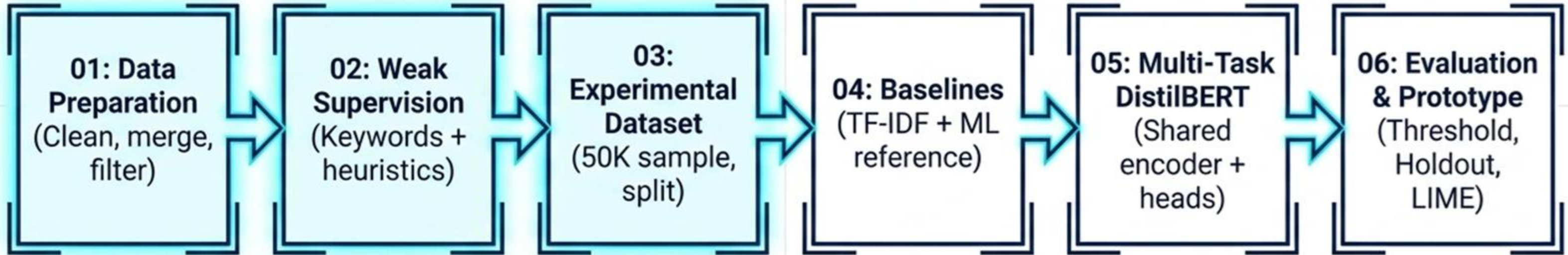
Email text → multiple categories + one priority class.

Comparative Analysis of Selected Studies

Research Direction	Selected Studies	Main Limitation
Classification/NLP	Fareed et al., 2025	Focuses on spam, not joint multi-label + priority.
Transformer Models	Jain et al., 2024	Contextual architecture not integrated with priority.
Priority/Urgency	Hirpara & Polara, 2025; Chuttur & Parianan, 2022	Isolated from multi-intent classification.
XAI Integration	Explainable AI (XAI)	Limited unified integration in one workflow.

The Gap: Fragmented tasks.
The Opportunity: A unified multi-label + priority + XAI framework.

End-to-End AI Project Lifecycle



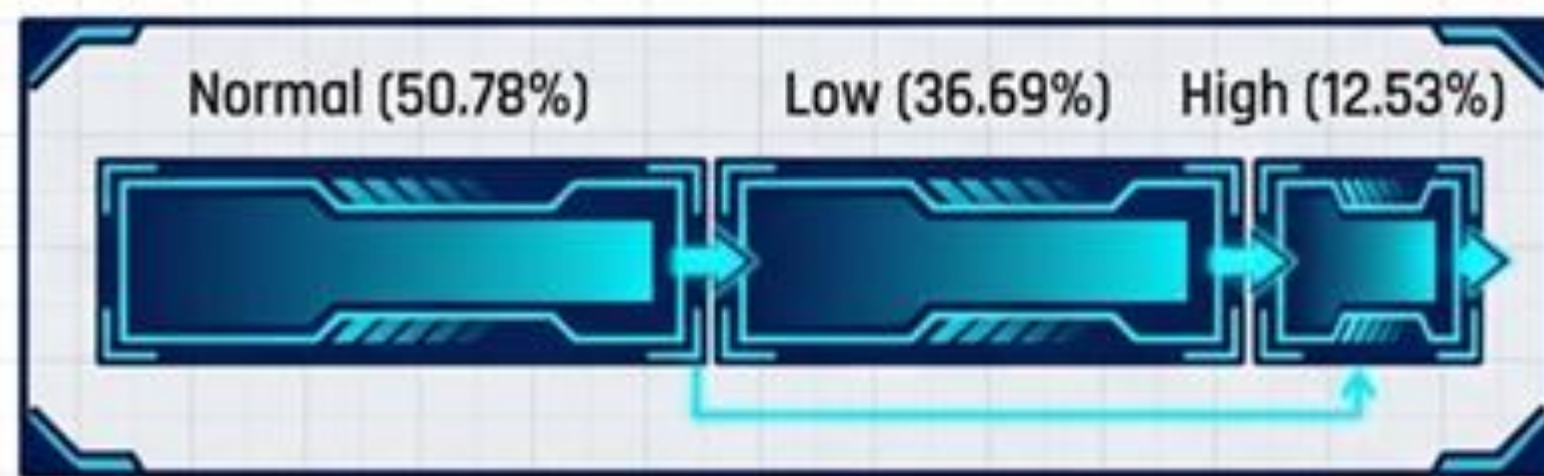
Research workflow: Data → Labels → Baselines → Proposed Model → Evaluation → Prototype.



Multi-Label Category Taxonomy



Single-Label Priority Taxonomy



Methodological Note: Generated via weak supervision (keyword dictionaries + heuristic rules) to serve as supervised training targets, not manual ground truth.

Strict Filtering Pipeline and Text Preprocessing

Strict Filtering Pipeline

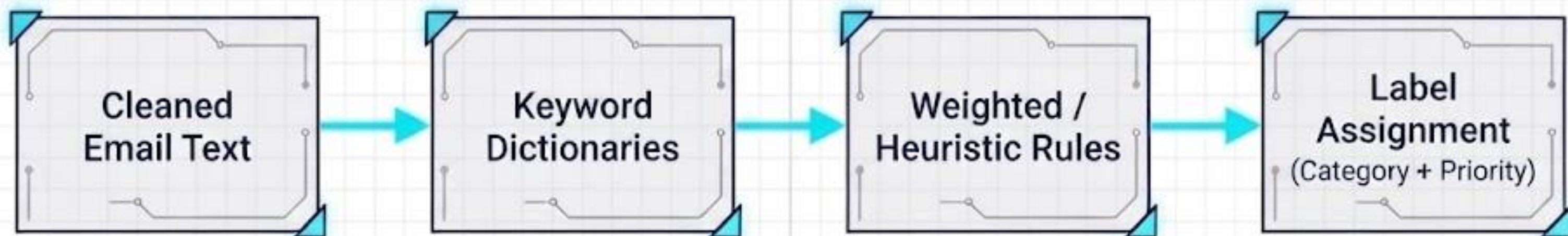


Text Preprocessing

- Subject + Body Merge
- Lowercase conversion
- URL removal
- Forward-header cleanup
- Special-character removal
- Whitespace normalization

Crucial Note: Empty-text removal performed; no stop-word removal or lemmatization is claimed.

Label Engineering via Weak Supervision



Category Taxonomy (Multi-label)

☐ Administration	☒ Legal
☒ Financial	☒ Market
☐ HR	☐ Other

Priority Taxonomy (Single-label)

☒ High
☐ Normal
☐ Low



Methodological Note: The generated labels served as supervised training targets. Programmatically generated labels – not manually annotated ground truth.

190,280 Cleaned Emails



**50,000 Final Experimental Sample
(Post-Stratified Sampling)**

Train: 35,000 (70%)
Purpose: Model fitting /
training.

Validation: 5,000 (10%)
Purpose: Model selection,
threshold optimization,
and monitoring.

Holdout: 10,000 (20%)
Purpose: Independent
final evaluation only.



Strict Protocol Rule: The Holdout Test Set was completely isolated. Not used for training, model selection, or threshold tuning. Final comparison relies on Holdout Test only.

Traditional ML Baseline Evaluation



Validation Results Comparison (Post-Model Selection)

Model	Category Macro F1	Priority Accuracy	Priority Macro F1
Logistic Regression	0.8751	0.8716	0.7900
Linear SVM	0.9267	0.8988	0.8565
Random Forest	0.6024	—	—

Takeaway: Linear SVM demonstrated highest validation performance and is selected as the strongest traditional baseline reference.

Note: These are validation results, not the final holdout comparison.

Cleaned Email
(Subject + Body)

Tokenizer

Shared DistilBERT Encoder
(Contextual representation)

Category Head
(Linear + Sigmoid • 6 labels)

Priority Head
(Linear + Softmax • 3 classes)

$$L_{total} = L_{category} + L_{priority}$$

Innovation Callout:

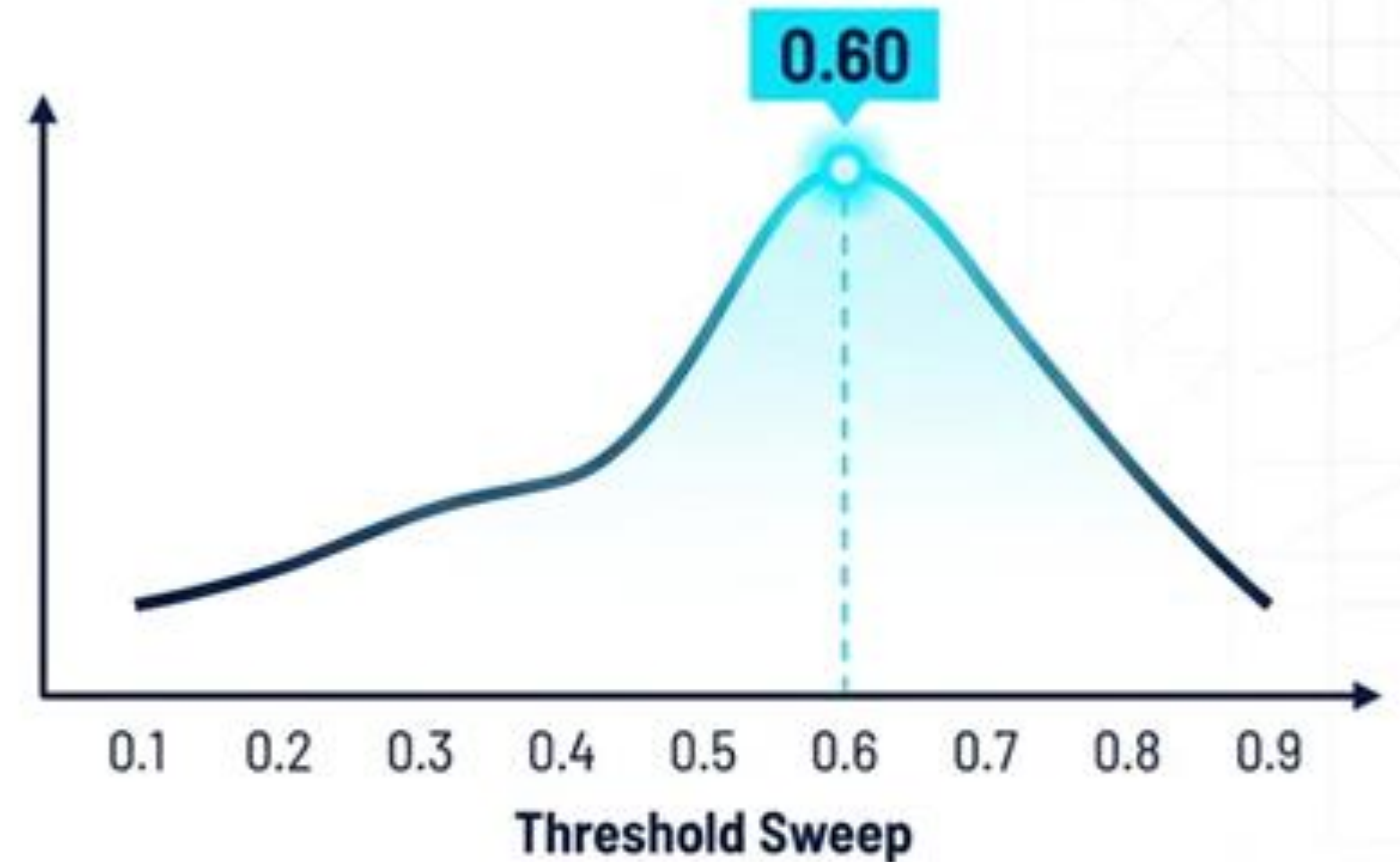
Key Innovation: A single shared contextual representation predicting multiple labels (multi-label classification) and priority (multi-class priority) simultaneously.

Training Configuration and Threshold Optimization Strategy

Parameter Dashboard

Optimizer:	AdamW
LR:	2e-5
Batch:	16
Epochs:	5
Loss:	BCE (Category) + Cross-Entropy (Priority)

Note: Mixed precision used



Default threshold (0.50)

Validation Macro F1: 0.9677 |
Hamming Loss: 0.01713

Optimized threshold (0.60)

Validation Macro F1: 0.9681 |
Hamming Loss: 0.01710

Protocol: Threshold 0.60 locked before Holdout. Holdout test strictly isolated from tuning.

Final Independent Holdout Results (Evaluated exclusively on 10,000 independent unseen emails).

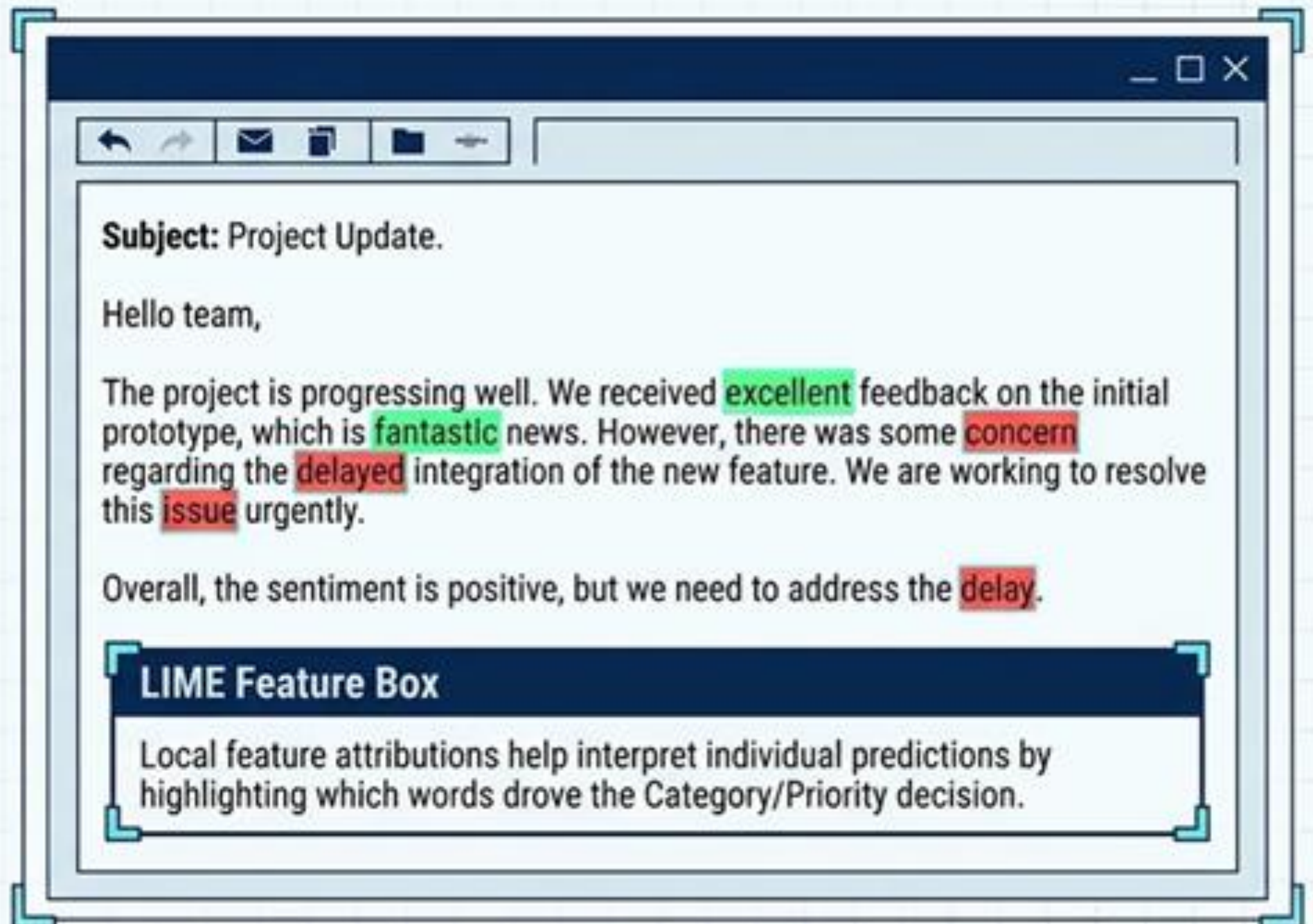
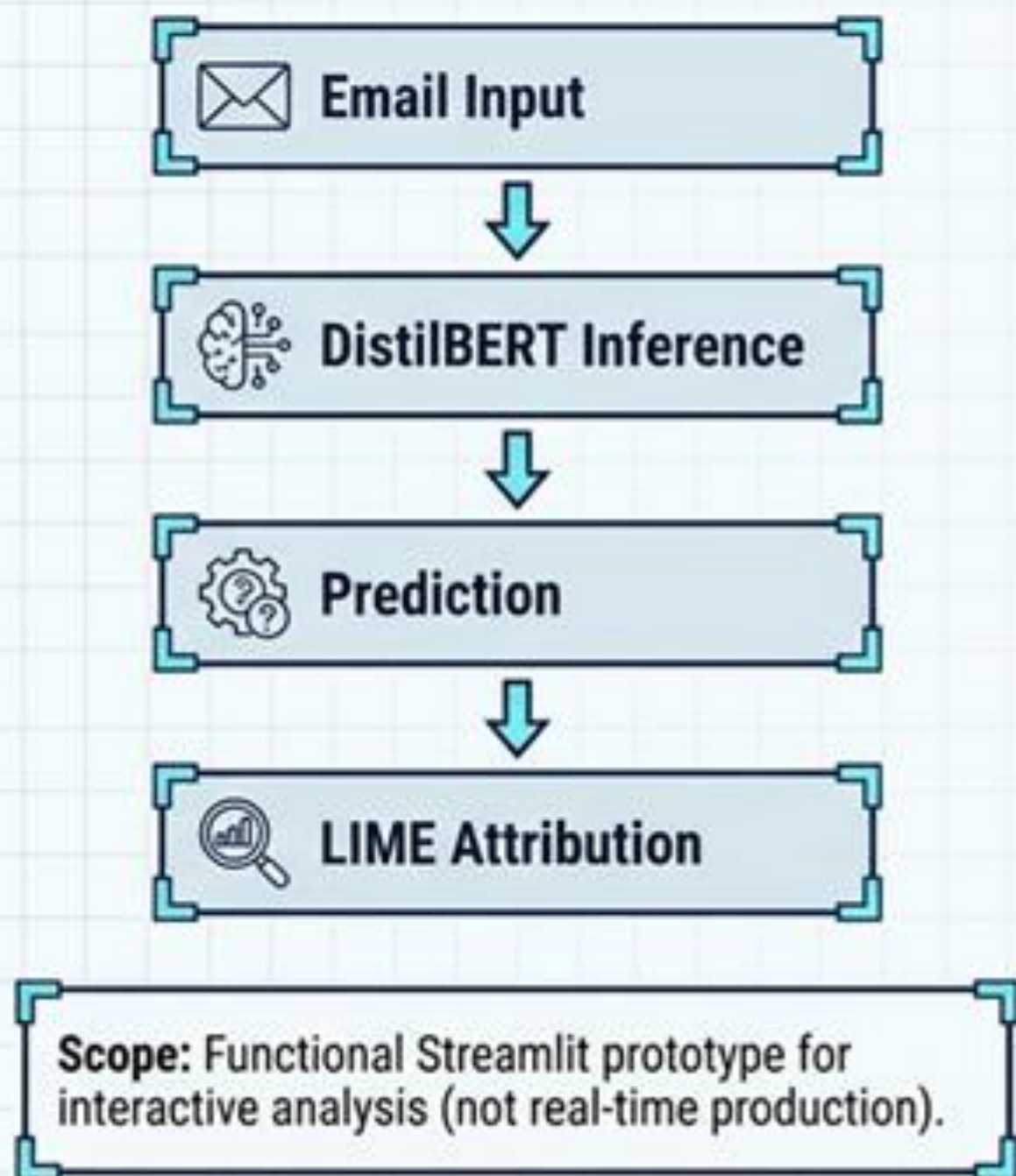
Metric	Linear SVM	DistilBERT	Improvement
Category Macro F1	0.9269	0.9672	+4.03 %
Priority Accuracy	0.9079	0.9529	+4.50 %
Priority Macro F1	0.8695	0.9327	+6.32 %
High-Priority Recall	0.6900	0.8627	+17.27 %

+17.27 %
High-Priority Recall

Significant leap in capturing critical high-priority items.

All four improvements are absolute percentage-point differences on the Holdout Set.

Explainable AI (XAI) Prototype and Interactive Dashboard



Contributions, Limitations, and Conclusion

Contributions



Unified multi-task framework



Weak-supervision pipeline



Optimized classification



XAI prototype

Limitations



Label noise from weak supervision



Compute cost vs linear baselines



Single historical dataset

Future Work



Human-in-the-loop validation



Multilingual support



Enterprise metadata integration

Final Verdict: Multi-Task DistilBERT successfully unified classification and priority prediction, **significantly outperforming baselines**