**Document Classification System**

A comprehensive guide for building an intelligent document classification system that can automatically categorize business documents by department and priority level.

**Overview**

This system helps organizations automatically classify and prioritize internal documents based on content analysis, keyword detection, and machine learning techniques. It supports classification across 12 common business departments and 3 priority levels.

**Supported Departments**

1. **Human Resources (HR)**
2. **Finance & Accounting**
3. **Legal**
4. **Sales**
5. **Marketing**
6. **IT (Information Technology)**
7. **Operations**
8. **Customer Support**
9. **Procurement / Purchase**
10. **Product / R&D**
11. **Administration**
12. **Executive / Management**

**Department Classification**

**1. Human Resources (HR)**

**Document Types:** Resumes, Offer Letters, Appointment Letters, Performance Reviews, Leave Applications, Training Materials

**Keywords:** employee ID, hiring, appraisal, benefits, recruitment, onboarding, resignation, PTO, attendance, HR policy

**2. Finance & Accounting**

**Document Types:** Invoices, Receipts, Balance Sheets, Profit & Loss Statements, Tax Filings, Salary Slips, Audit Reports

**Keywords:** invoice, payment, accounts payable, ledger, tax, balance sheet, fiscal, payroll, expenses, revenue, debit, credit

**3. Legal**

**Document Types:** Contracts, NDAs, MOUs, Compliance Reports, Policies, Lawsuits, Legal Notices

**Keywords:** non-disclosure, contract, agreement, terms, regulation, compliance, clause, legal, dispute, jurisdiction, breach

**4. Sales**

**Document Types:** Sales Reports, Proposals, Quotations, Purchase Orders, CRM Records

**Keywords:** sales target, lead, quotation, conversion, pipeline, customer, deal, revenue, proposal, client

**5. Marketing**

**Document Types:** Campaign Plans, Social Media Reports, SEO Analysis, Branding Guides, Event Planning Docs

**Keywords:** campaign, branding, SEO, email blast, content, engagement, target audience, lead generation, ad spend

**6. IT (Information Technology)**

**Document Types:** Incident Reports, IT Policies, User Guides, System Logs, Network Diagrams, SLA Reports

**Keywords:** server, network, incident, troubleshooting, firewall, access control, login, cybersecurity, SLA, IT support

**7. Operations**

**Document Types:** Process Manuals, Daily Logs, Maintenance Reports, SOPs, Supply Chain Docs

**Keywords:** logistics, supply, workflow, daily operations, SOP, inventory, maintenance, efficiency

**8. Customer Support**

**Document Types:** Support Tickets, Chat Logs, Feedback Reports, Service Reports

**Keywords:** ticket, customer issue, response time, escalation, helpdesk, satisfaction, support team, client query

**9. Procurement / Purchase**

**Document Types:** Purchase Orders, Vendor Agreements, RFQs, Bills, Delivery Notes

**Keywords:** purchase order, vendor, quotation, invoice, RFQ, delivery, procure, supplier, inventory

**10. Product / R&D**

**Document Types:** Product Specs, Design Docs, Testing Reports, Bug Reports, Research Papers

**Keywords:** feature, testing, prototype, bug, release, version, specification, roadmap, R&D

**11. Administration**

**Document Types:** Facility Requests, Asset Allocation, General Notices, Office Supplies Requisitions

**Keywords:** facility, stationery, asset, building maintenance, admin, general request, supplies

**12. Executive / Management**

**Document Types:** Strategy Documents, Board Meeting Notes, Annual Reports, Vision Statements

**Keywords:** strategy, KPI, vision, mission, goals, board, agenda, quarterly review, annual report

**Priority Classification**

**High Priority (Time-sensitive, urgent, requires immediate action)**

| **Category** | **Keywords/Phrases** |
| --- | --- |
| **Deadlines** | by EOD, by end of day, by today, asap, urgent, immediate, within 24 hours, deadline today, due today, respond by, reply immediately |
| **Action Requests** | action required, requires immediate attention, please review urgently, high priority, critical issue, resolve now |
| **Escalations/Issues** | escalated, service disruption, breach, incident, system down, customer complaint, payment failed |
| **Meetings/Events** | today's meeting, final review, must attend, confirmation needed |

**Medium Priority (Important but not urgent — typically this week or within a few days)**

| **Category** | **Keywords/Phrases** |
| --- | --- |
| **Follow-ups** | reminder, follow up, this week, pending, awaiting response, check status, update needed |
| **Upcoming Deadlines** | by tomorrow, due in 2 days, schedule by, before Friday, complete by, ETA |
| **Meetings** | scheduled for, calendar invite, tentative, planned discussion, agenda |
| **Tasks** | work in progress, assigned, need update, submit by, to be reviewed |

**Low Priority (Informational, long-term, or low urgency)**

| **Category** | **Keywords/Phrases** |
| --- | --- |
| **FYI/Reference** | for your information, no action needed, for record, just sharing, reference document, read only, optional |
| **Long-Term** | next quarter, next month, future release, roadmap, tentative plan, long-term goal, backlog item |
| **General Updates** | weekly summary, monthly report, feedback, draft version, notes, not urgent |

**Implementation Approaches**

**1. Rule-Based Classification**

Start with a baseline keyword matching system using the provided keywords for each department and priority level.

**2. Machine Learning Models**

Enhance the rule-based system with:

* **TF-IDF + Logistic Regression/SVM**
* **BERT-based classifier** (fine-tuned for your document corpus)
* **Spacy or Transformers** for Named Entity Recognition and similarity-based classification

**3. Multi-Label Classification**

Train a classifier to predict both department and priority simultaneously for more efficient processing.

**Usage Guidelines**

1. **Keyword Matching**: Build a regex or keyword matching system to assign priority\_level: high | medium | low
2. **NLP Enhancement**: Combine keyword matching with NLP models or TF-IDF ranking for dynamic results
3. **Multi-Label Training**: Train models to predict both department and priority labels simultaneously
4. **Continuous Learning**: Regularly update keywords and retrain models based on new document types

**Getting Started**

1. Implement the keyword-based classifier as a baseline
2. Collect and label your organization's documents
3. Train ML models on your specific document corpus
4. Integrate with your document management system
5. Monitor and improve classification accuracy over time

**Best Practices**

* Start with rule-based classification for quick implementation
* Gradually introduce ML models for improved accuracy
* Regularly update keyword lists based on organizational changes
* Use feedback loops to continuously improve classification performance
* Consider document context and organizational structure when fine-tuning