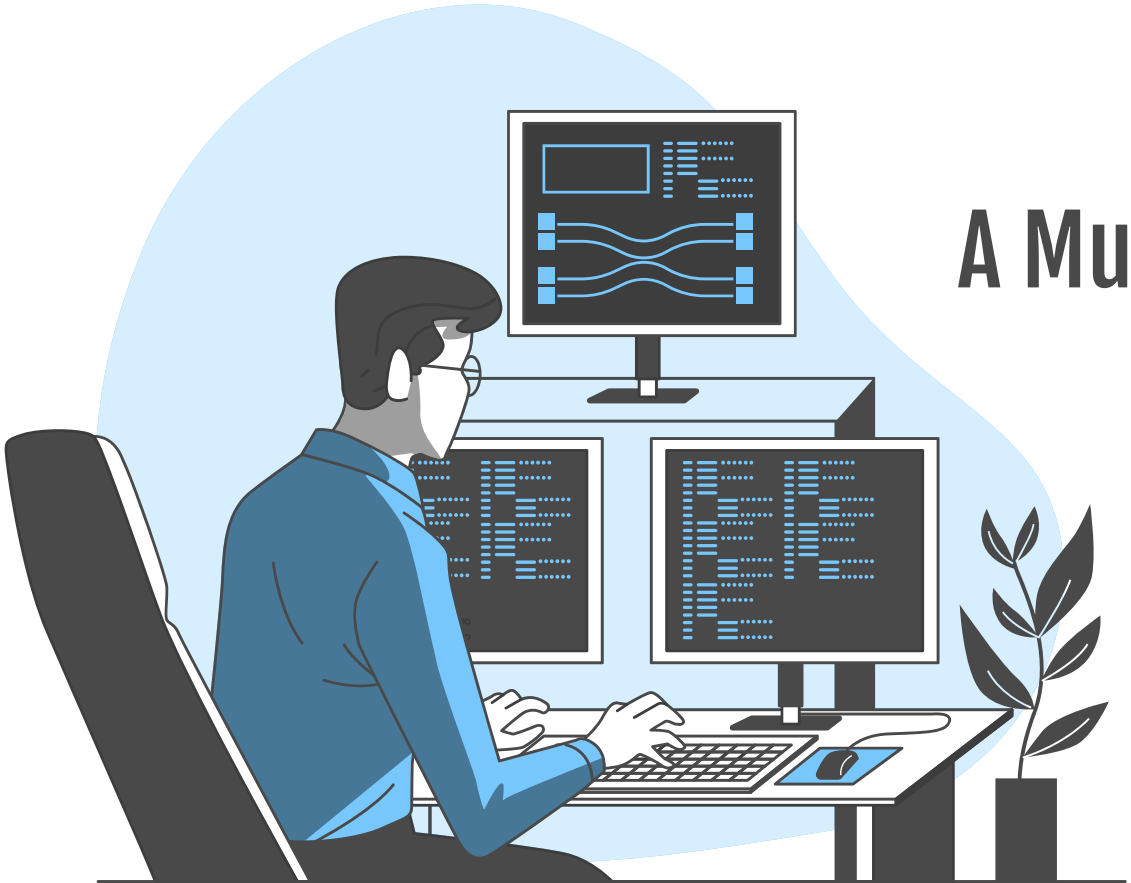
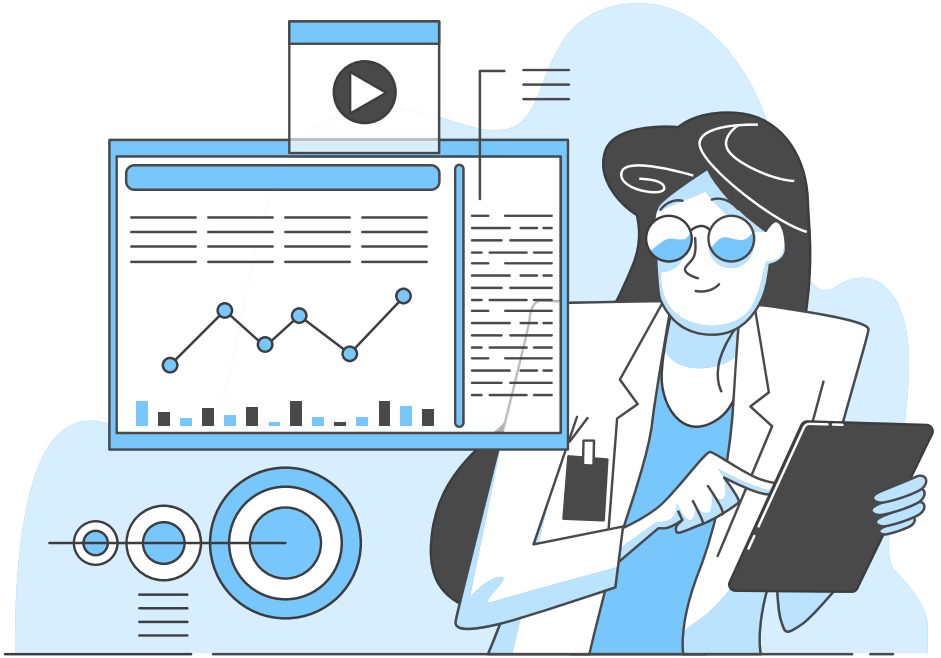


A Multi-Page Document Classification Tool

Target Customers:
Small to Large Businesses

Team 5:
Sunit Carpenter
Joseph Issa
Amber Chen
Sissie Cui

Table of Contents



Business Problem

What's the Problem?

Many businesses have large amounts of unstructured files in their file repositories where **files are added constantly but never organized or archived properly**

Decrease Utility Efficiency

Creates duplication of files

Decrease Productivity

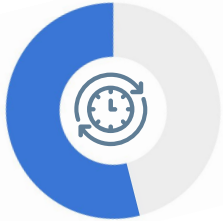
Makes it difficult to search for information

Increase Cost

Increase file storage expenses

Business Opportunity

54%



report **wasting time searching** for much-needed files in cluttered online filing systems

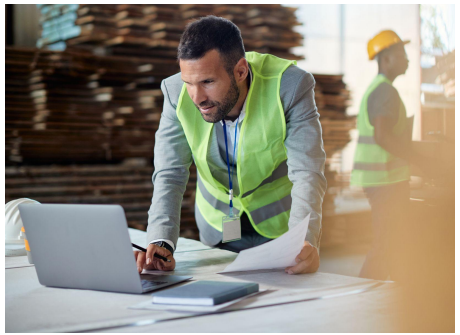
57%

rank **being able to quickly find the files and document they need** as a top three problems to solve

Source: Wakefield Research Survey n = 1000 U.S. office professionals

Customer Persona

Meet Our Customer: Thomas Hill



Age: 40

Occupation: Project Manager

Industry: Commercial Construction

Personality

- He is an efficiency maximizer. If there is a better way to do things, he will find it.

Pain Points

- He receives many invoices and receipts, but **does not have time to categorize files** so it is hard to retrieve files when needed
- One project file repository could be used by multiple project managers
- No one likes organizing files so the **share folder is cluttered with duplicated and outdated files**

Document Types He Handles

Desired Solutions

1. Automatically classify documents
2. Automatically move documents to the destination folders
3. Automatically archive outdated files based on customized rules

Our Solution

01

Create a desktop app that **automatically categorizes documents** in file repositories into the common document types

02

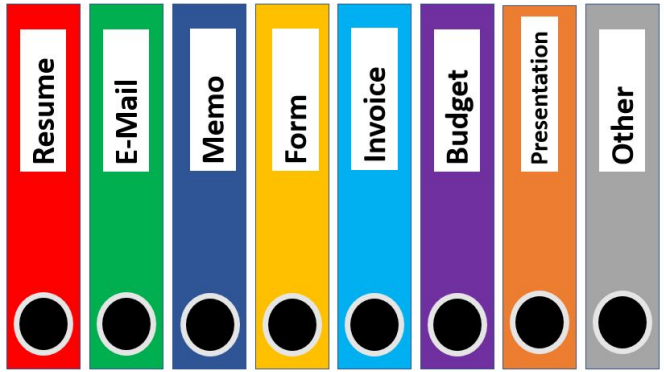
Summarize the content to save time in reading the documents

03

Enable users **to customize the classification model** and archive rules

04

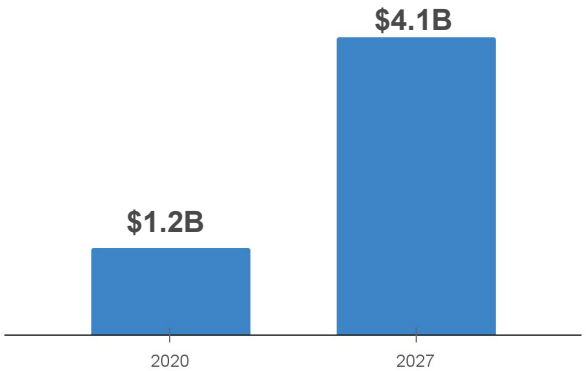
Automatically move files to destination folders based on document types and archive outdated files



Market Opportunity & Competition

Growing Global Demand

In 2020, the market size of intelligent document processing was at $1.2 billion, which is projected to reach **$4.1 billion** by 2027



Harnessing Unstructured Data

80% of worldwide data will be **unstructured** by 2025

Fragmented Market Presents Opportunity for Market Entry

Most of the current players are small organizations that heavily focus on data extraction

DOCSUMO

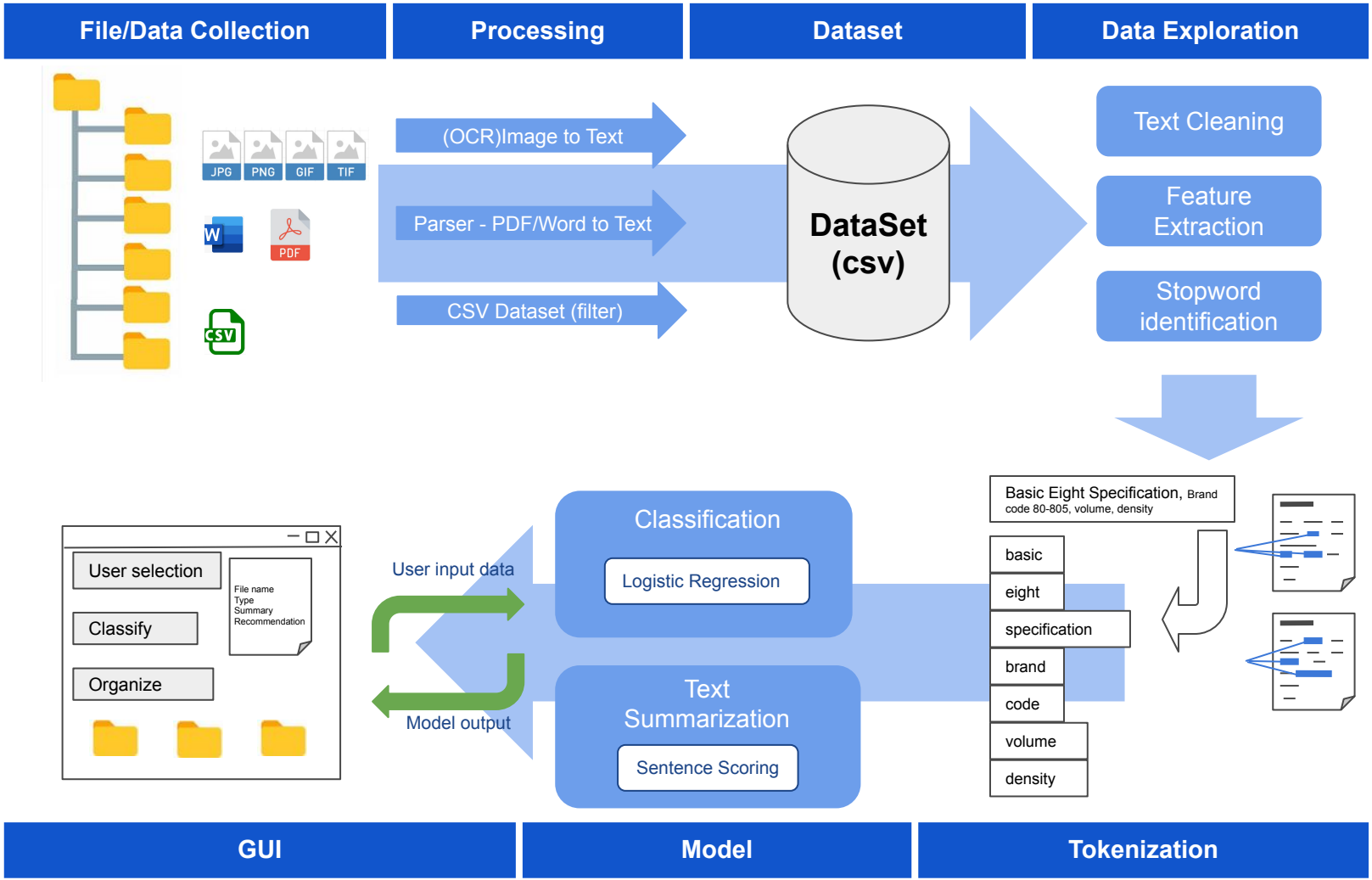
GROUPDOCS

SuperAnnotate

Source: [Competitive Landscape: Intelligent Document Processing Platform Providers (gartner.com)]

Our Solution

PageWise Architecture



Data Collection and Exploration

200k+ documents

Collected from multiple sources (Kaggle, Philip Morris, etc.)

3 File Formats

PDF, word (.docx), image (.jpg, .tif)

Scope Selection

10 most common document classes for small to large businesses

Balanced Sample Size

20k+ documents for every document class

Vector Representation

A Bag of Clean Words

Removed numbers, punctuations and stop words

Tokenization

Built a vocabulary of words in the corpus

Text Vectorization

Counted the occurrences of words presented in each document

Performance Boosted by Feature Engineering

Base Model

80% Accuracy
by logistic regression

3 Engineered Features

Document Length
total number of words

Numeric text ratio
Numerical proportion of the doc

Word uniqueness ratio
Number of unique words by total length

Performance Boost

2 ~ 5%
increase in accuracy

Model Evaluation

Logistic Regression

Score: 84%

Pros

- Best accuracy
- Smallest model file size
- High training efficiency
- Fastest for GUI responsiveness

Cons

- Unable to solve nonlinear problems

Naive Bayes

Score: 71%

Pros

- High training efficiency

Cons

- Lack of predictability for continuous numerical values

Random Forest

Score: 77%

Pros

- Amplify predictivity probabilities through multiple decision tree

Cons

- Long training time

BERT

Score: 70%

Pros

- Stronger awareness of the context of each text it analyzes

Cons

- High computational expense
- High memory requirements

Model Selection

Logistic Regression with CountVectorizer

Performance Testing

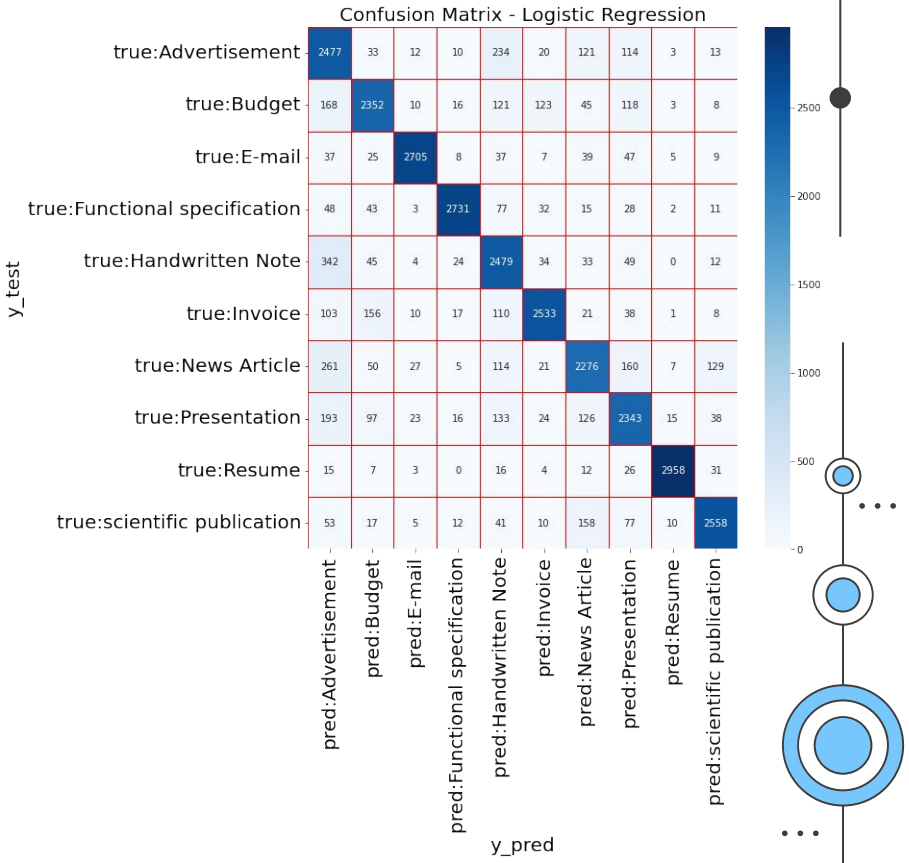
In-sample Validation Test

84% accuracy

Out-of-sample Scenario Test

Client: Career Coach

90% accuracy



Product Demo

GUI APP

[Product Website]

Potential Monetization Opportunities

Software as a Product

- One-time purchase license
- Subscription-based license
- Enterprise license

SAAS + Consulting Services

- Subscription-base license
- Enterprise license
- Tech implementation services
- Strategy consulting services

An Integration with OS (Native App Store App)

- One-time purchase license
- Subscription-based license
- Enterprise license
- Acquisition opportunity

PageWise Case Study

Client: Department of Agriculture

Problem:

- File repositories needs to be organized
- Unique file types:

Agroecosystems	Plants & Crops
Agricultural Economics	Animals & Livestock
Bioenergy	Food & Nutrition

Solution: PageWise (SAAS + Consulting Services)

- Collect sample files with labels
- Customize the classification models
- Select the best model that meets the client's needs
- Deploy the customized app on client site
- Provide ongoing technical support

Future Development

Technical & Time Constraints

App UI responsiveness constrained by...

Model File Size

Logistic Regression	230 MB
Naive Bayes	6 GB
Random Forest	1.2 GB

Python GUI Framework

- Lower performance when running a large-size model compared to .Net native framework

To deliver the MVP within 14 weeks...

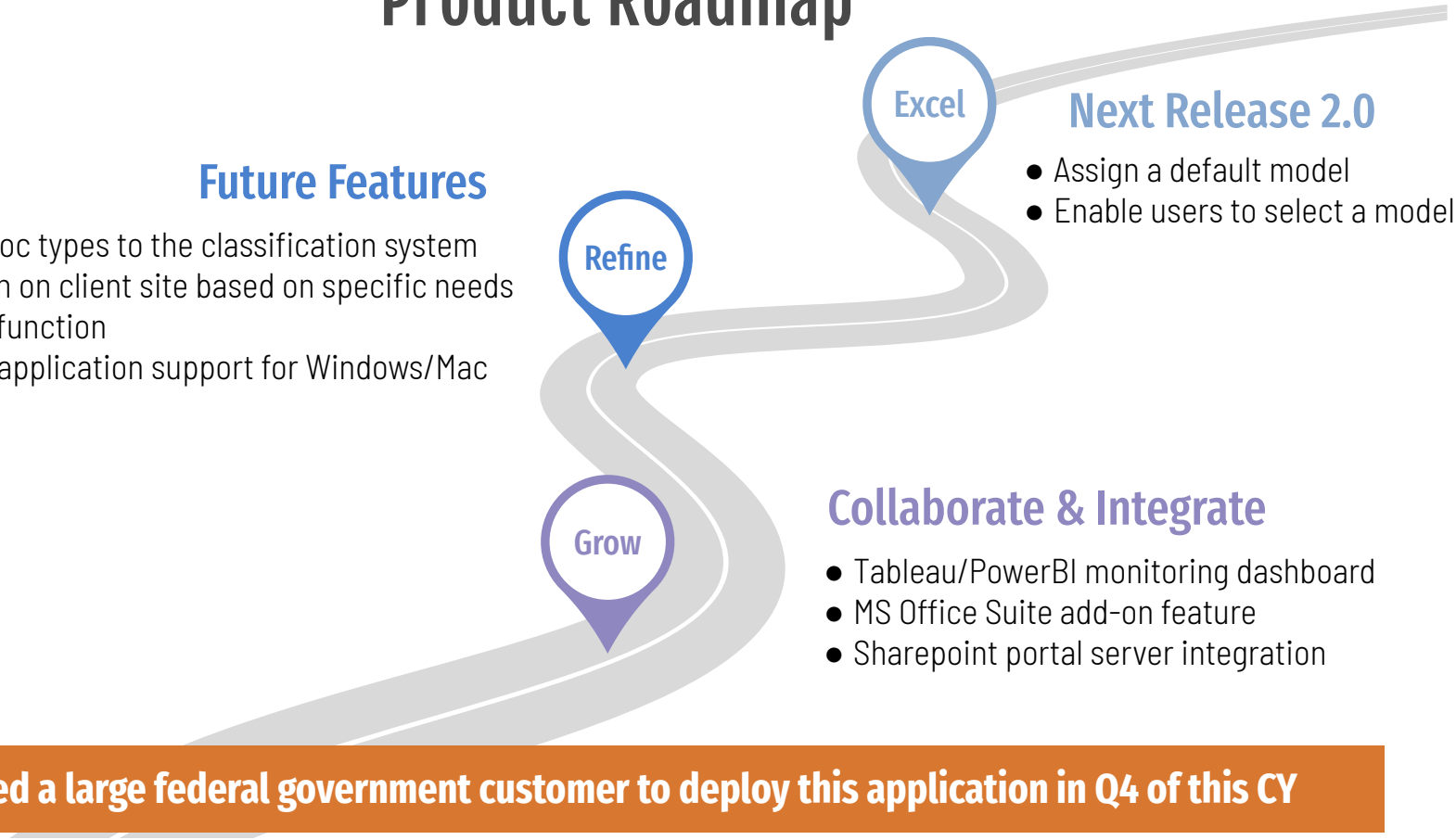
Simplified Summarization Model

- Sentence scoring was selected over BERT
- Compromised with the simpler model but average performance

Limited Scope of File Types

- Capped to the 10 most common types for a general business scenario

Product Roadmap



Next Release 2.0

- Assign a default model
- Enable users to select a model

Future Features

- Additional doc types to the classification system
- Auto re-train on client site based on specific needs
- Watermark function
- .Net Native application support for Windows/Mac

Collaborate & Integrate

- Tableau/PowerBI monitoring dashboard
- MS Office Suite add-on feature
- Sharepoint portal server integration

Secured a large federal government customer to deploy this application in Q4 of this CY

Thank you!

Appendix

References

Vijay Kumar, G., Yadav, A., Vishnupriya, B., Naga Lahari, M., Smriti, J., & Samved Reddy, D. (2021). Text summarizing using NLP. *Recent Trends in Intensive Computing*. https://doi.org/10.3233/apc210179

Python guis for humans. PySimpleGUI. (n.d.). Retrieved August 4, 2022, from https://www.pysimplegui.org/en/latest/

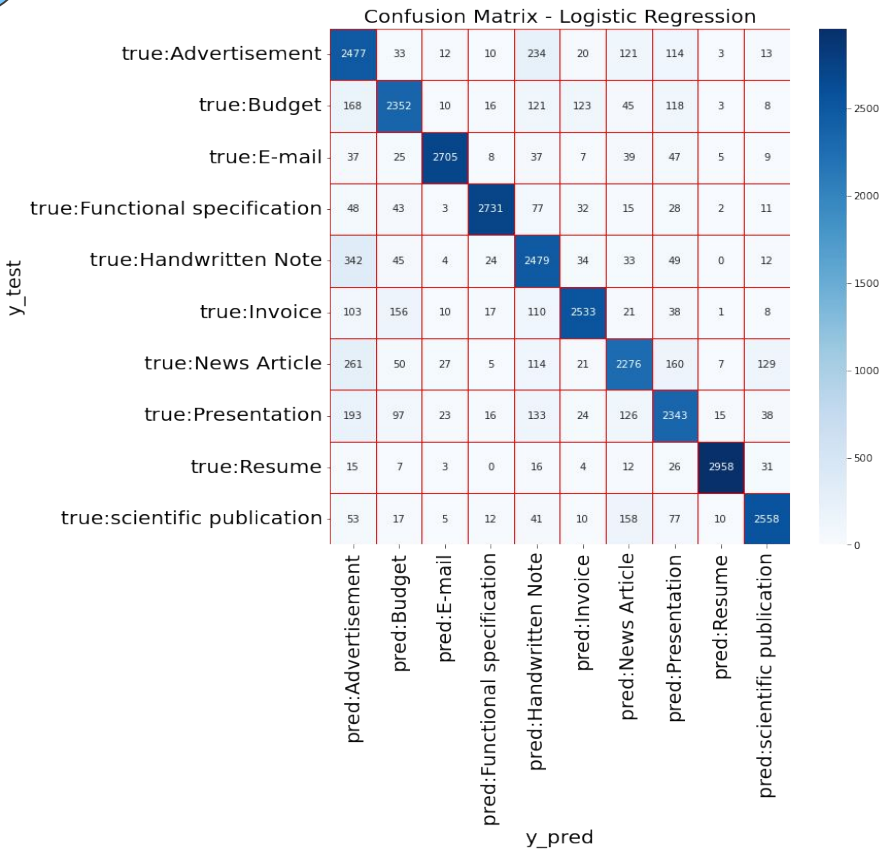
Basarkar, A. (2017). Document classification using machine learning. *DOCUMENT CLASSIFICATION USING MACHINE LEARNING*. https://doi.org/10.31979/etd.6jmu-9xdt

Wagh, V., Khandve, S., Joshi, I., Wani, A., Kale, G., & Joshi, R. (2021, November 1). *Comparative study of long document classification*. arXiv.org. Retrieved August 4, 2022, from https://arxiv.org/abs/2111.00702

The RVL-CDIP dataset. RVL-CDIP Dataset. (n.d.). Retrieved August 4, 2022, from https://www.cs.cmu.edu/~aharley/rvl-cdip/

Gartner_Inc. (n.d.). *Competitive landscape: Intelligent document processing platform providers*. Gartner. Retrieved August 4, 2022, from https://www.gartner.com/en/documents/4008008

Model Evaluation (LR)

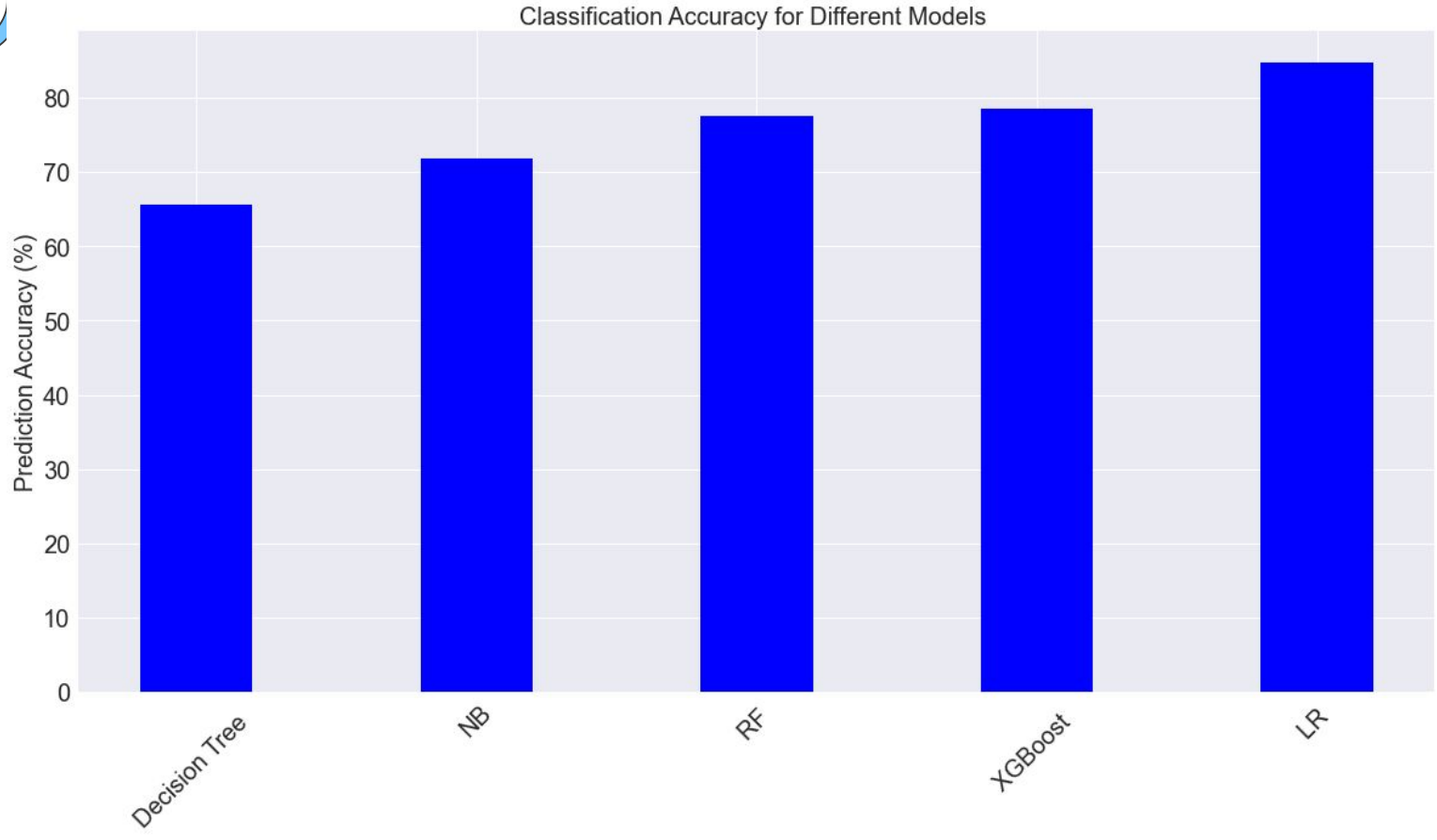


Prediction Accuracy

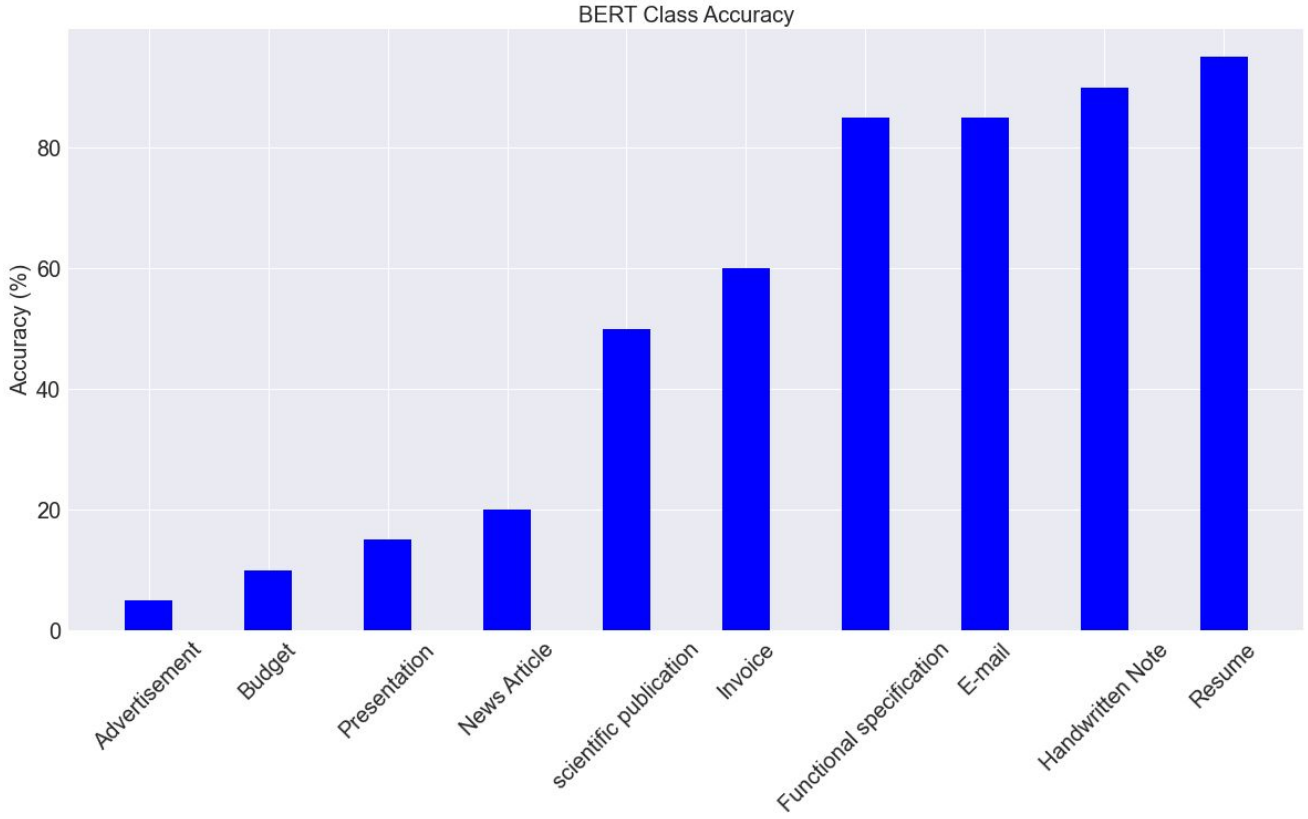
In-sample Dev test: 84%

Out-of-sample test: 90%

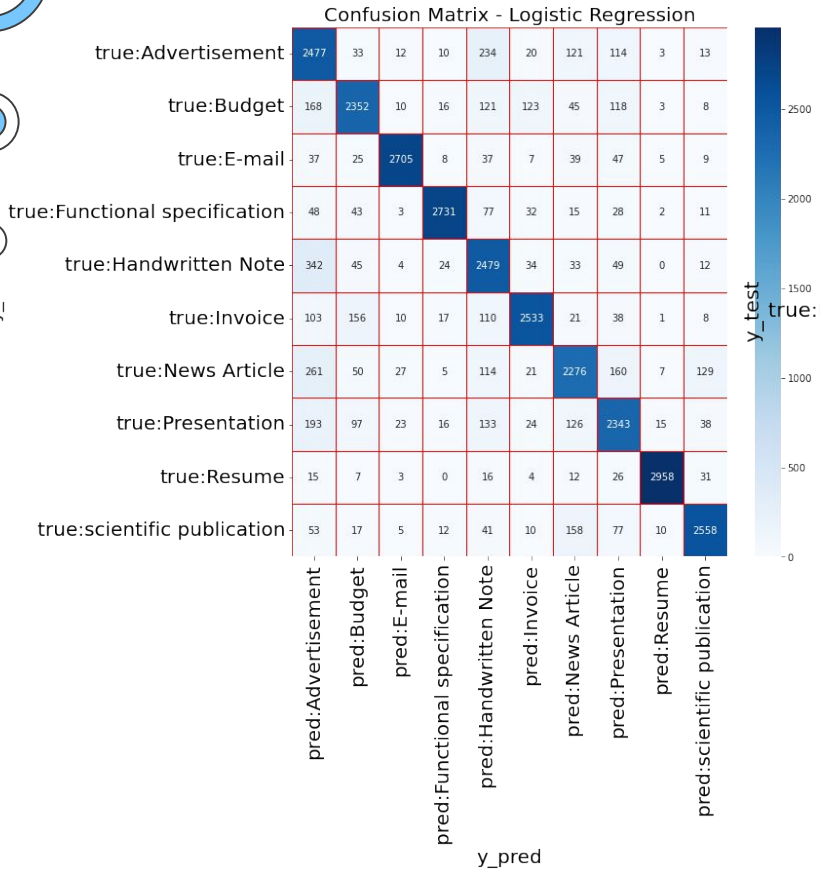
Model Evaluation

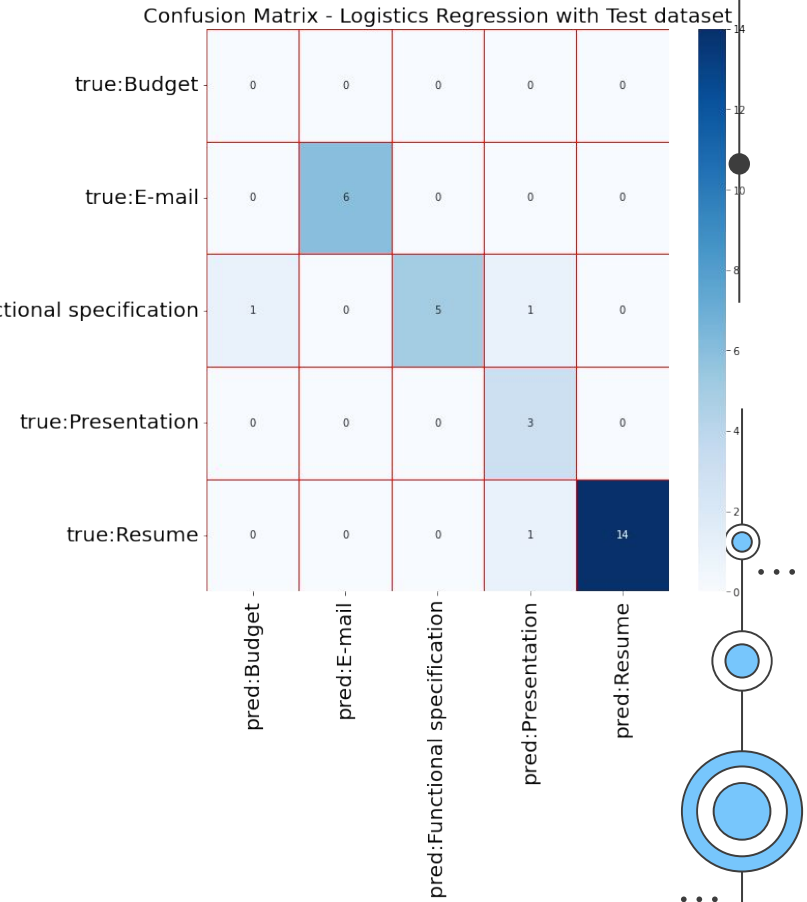


BERT Evaluation (Small Data Set)

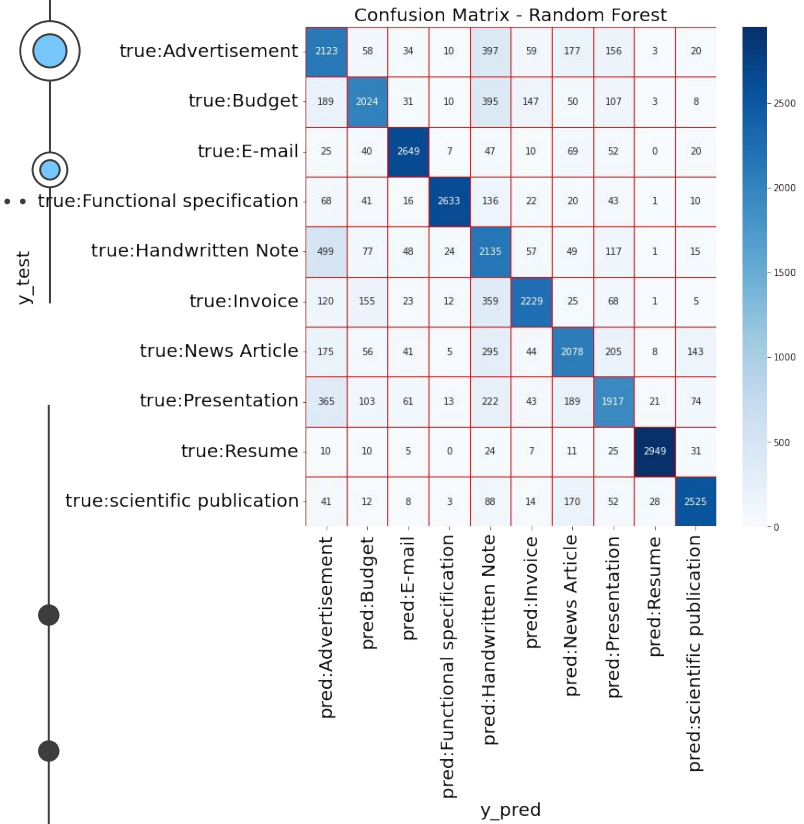


Model Evaluation (LR)





Module Evaluation (RF)



Module Evaluation(NB)