

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

01

Business Problem

02

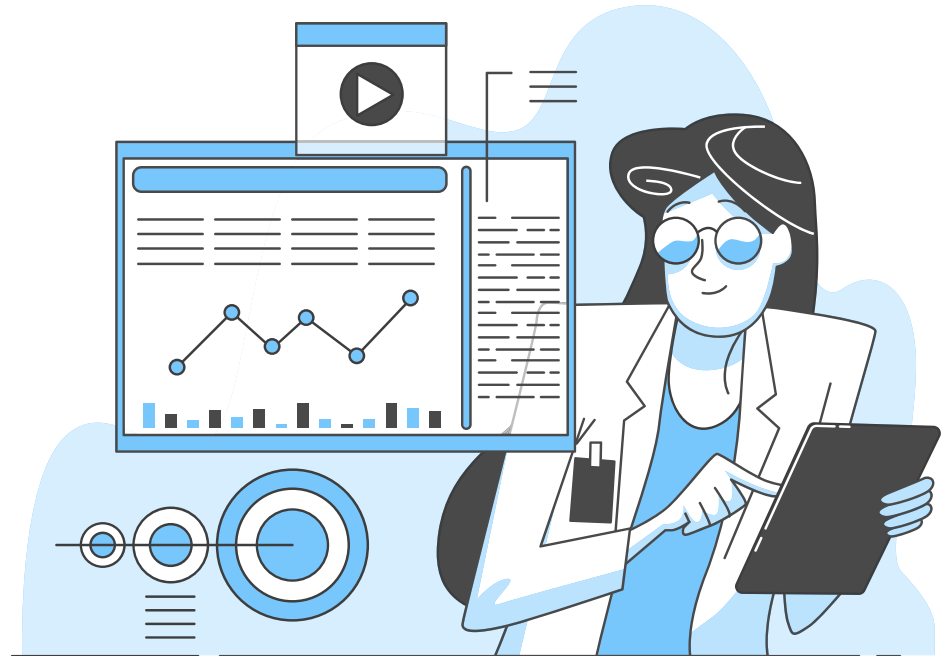
Our Solution

03

Product Demonstration

04

Future Development

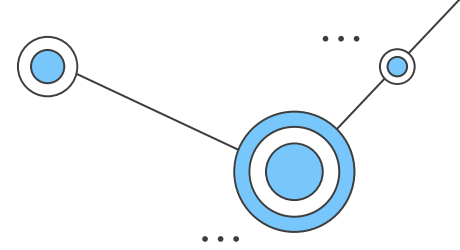




Business Problem



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**

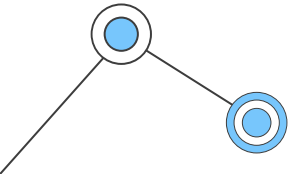
Creates duplication of files

**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

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- No one likes organizing files so the **share folder is cluttered with duplicated and outdated files**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

Invoices



Emails



Specifications



Desired Solutions

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- 3 Automatically archive outdated files based on customized rules

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

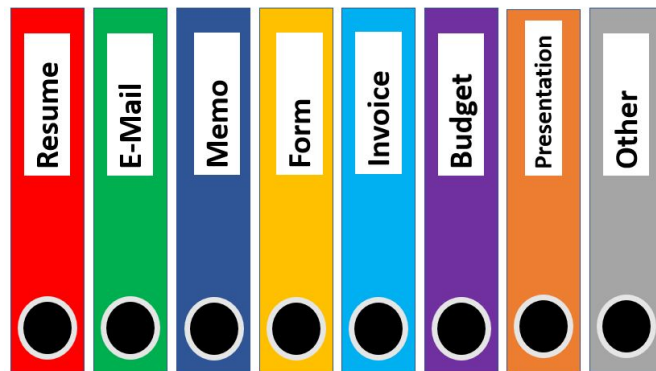
03
...

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

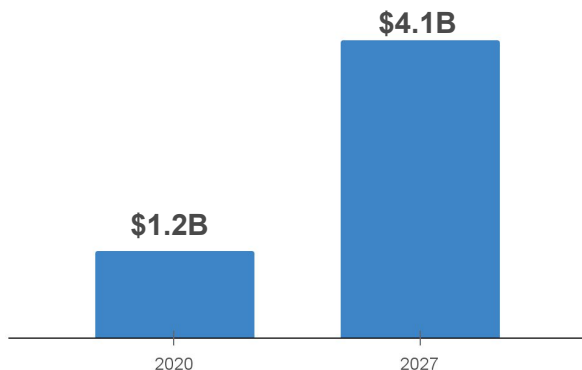
Our Solution



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



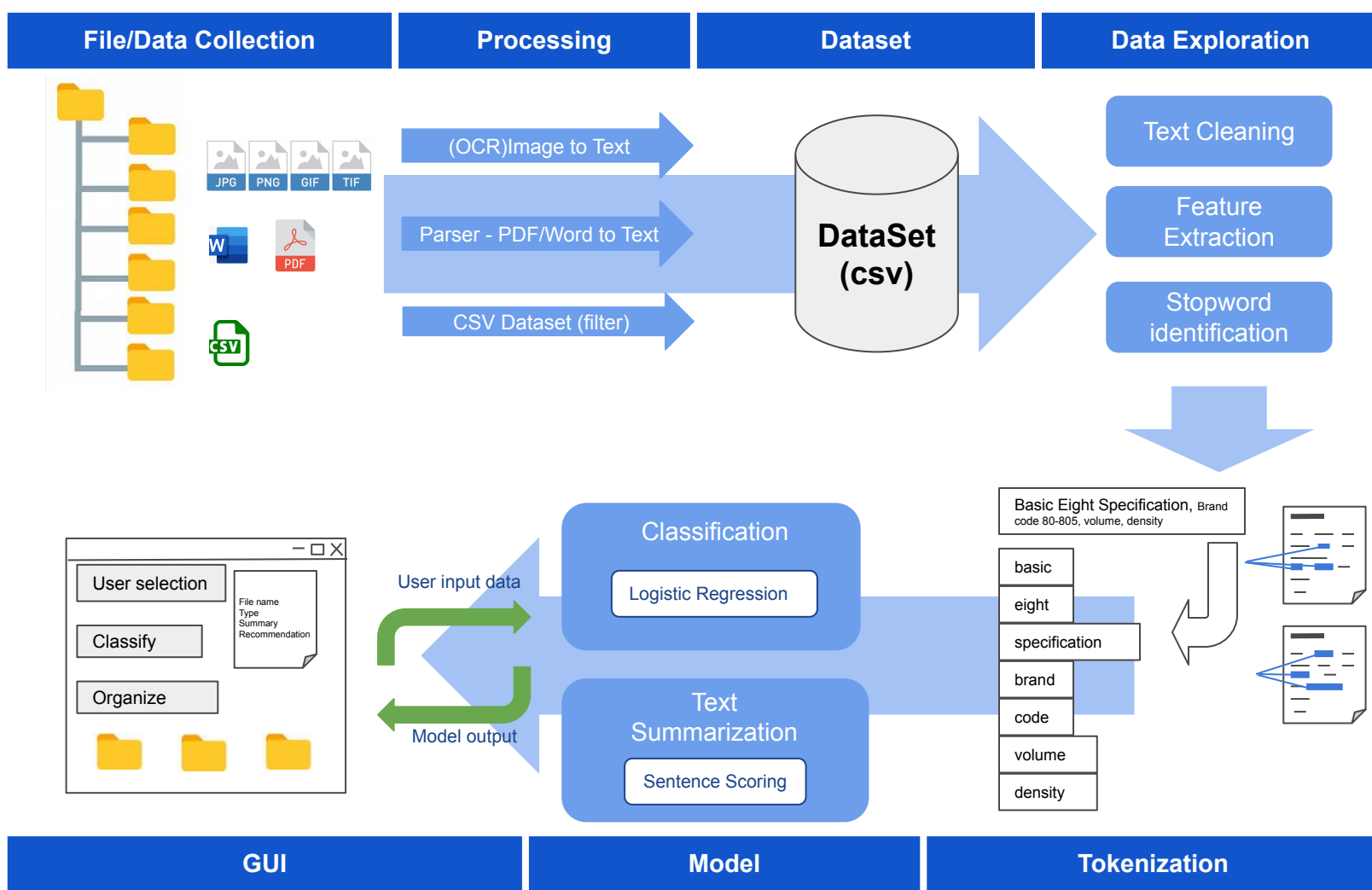


Our Solution



The image features a light blue, irregular blob shape in the center, which serves as a background for the text. Surrounding this central element is a decorative network diagram. This diagram consists of several blue circular nodes connected by thin black lines. Three of these nodes are larger and feature concentric circles, while the others are smaller. The nodes are positioned at the top right, bottom left, and bottom center of the frame, with some lines extending towards the edges, suggesting a larger, unseen network.

PageWise Architecture



Data Collection and Exploration



200k+ documents

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



**Vector
Representation**

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



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

3 Engineered Features



Base Model

80% Accuracy
by logistic regression

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

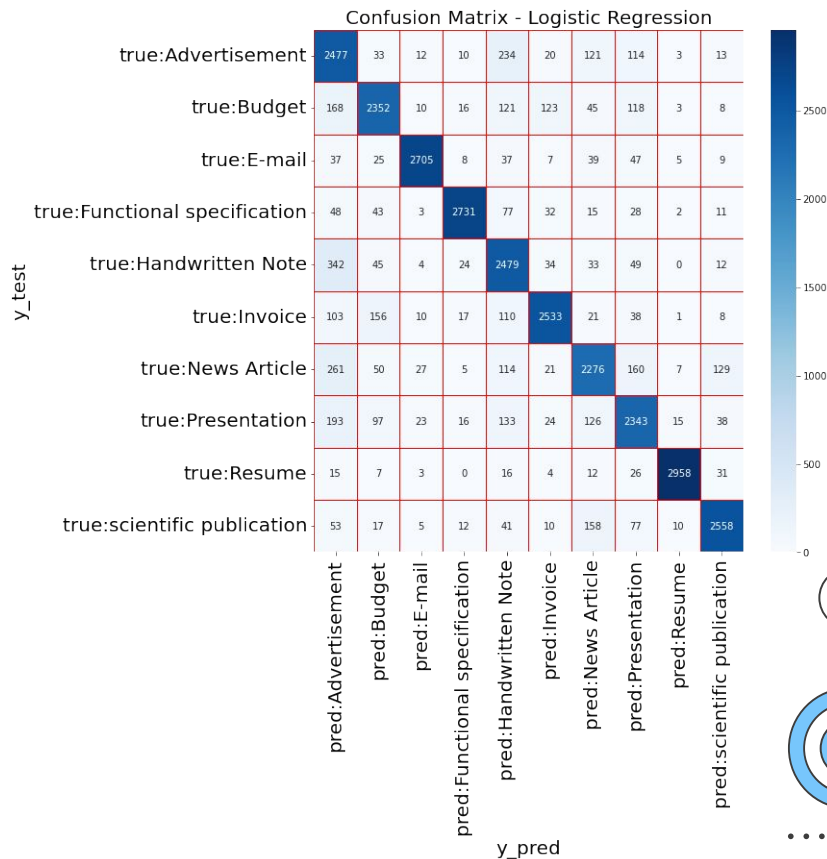
84% accuracy

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:

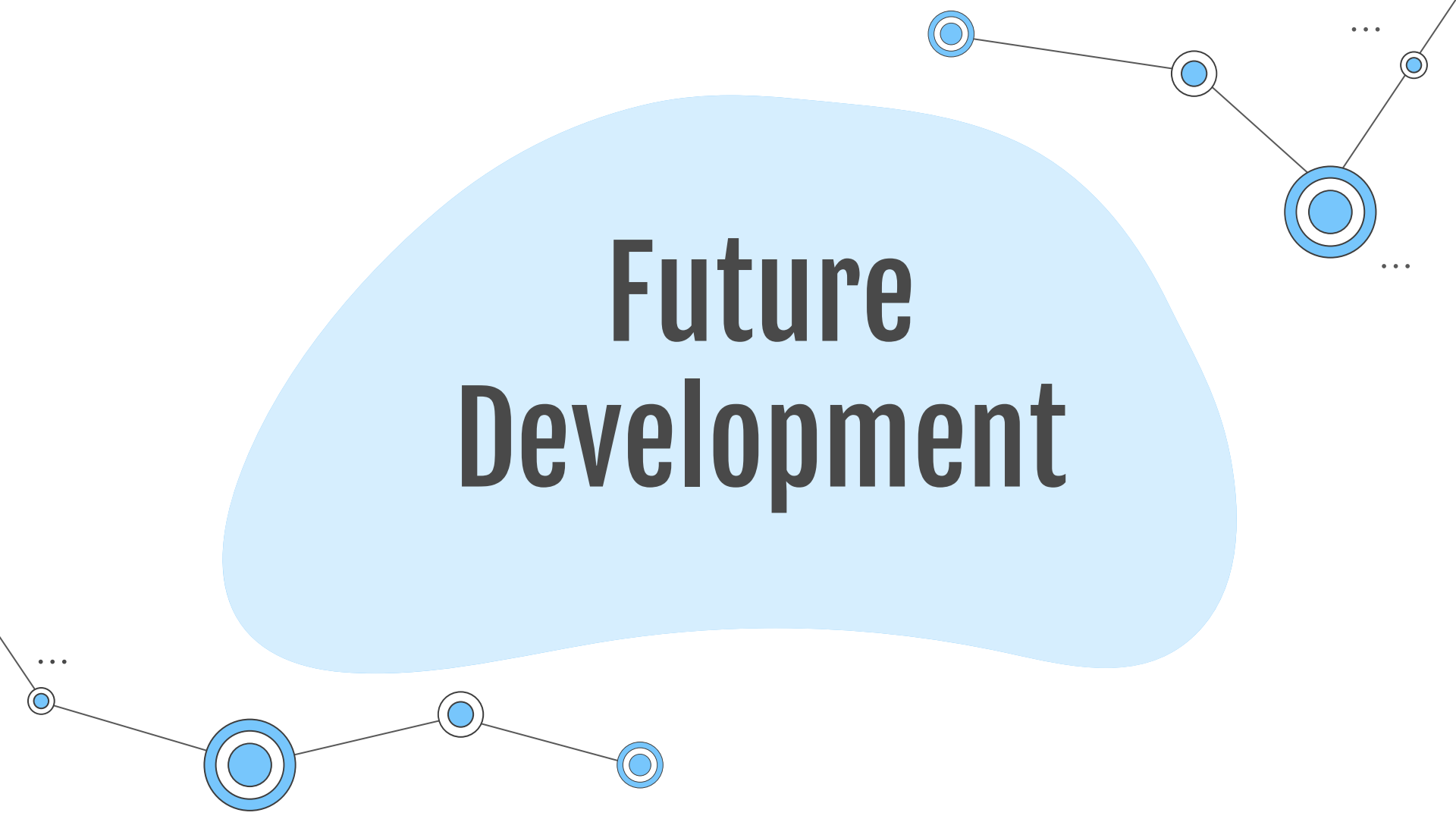
Agroecosystems	Plants & Crops
Agricultural Economics	Animals & Livestock
Bioenergy	Food & Nutrition



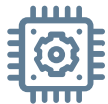
Solution: PageWise (SAAS + Consulting Services)

- 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

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

Refine

Grow

Excel

Next Release 2.0

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

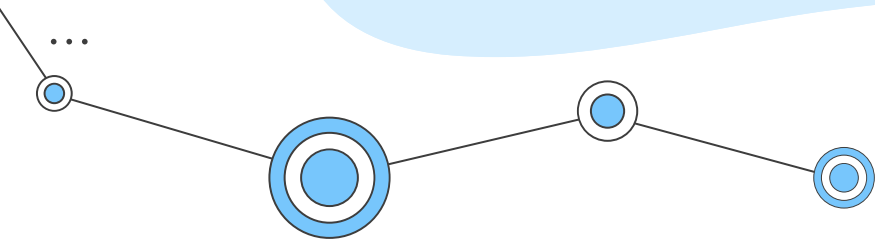
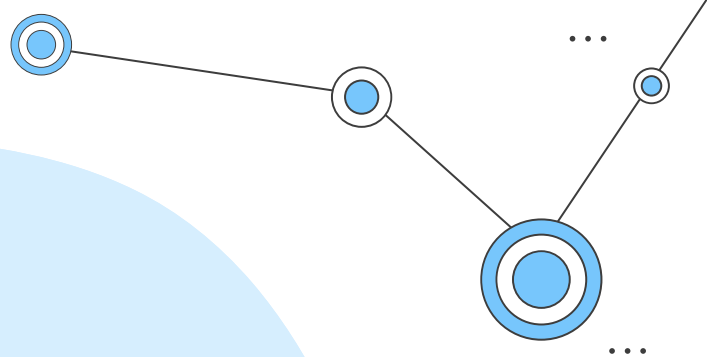
Collaborate & Integrate

- 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

A decorative network diagram is positioned in the background. It features several blue circular nodes of varying sizes, some with concentric circles, connected by thin black lines. The nodes are arranged in a non-linear fashion, with some branching out from a central point. Ellipses (...) are used to indicate that the network continues beyond the visible nodes.

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/>

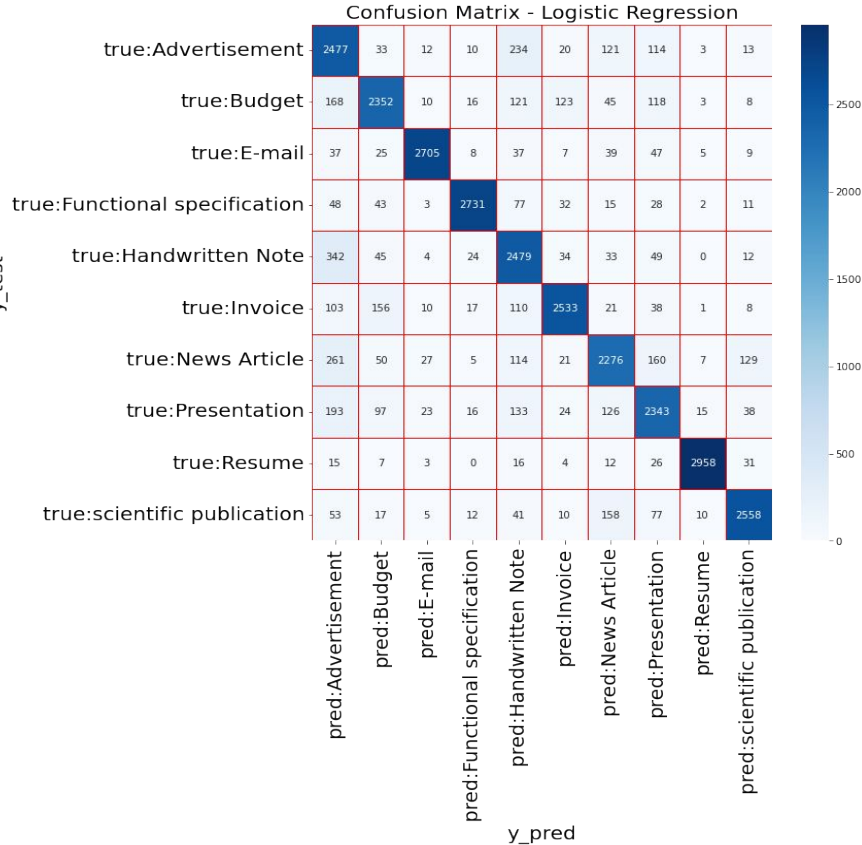
Basarkar, A. (2017). Document classification using machine learning. *DOCUMENT CLASSIFICATION USING MACHINE LEARNING*. <https://doi.org/10.31979/etd.6jmu-9xdt>

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)

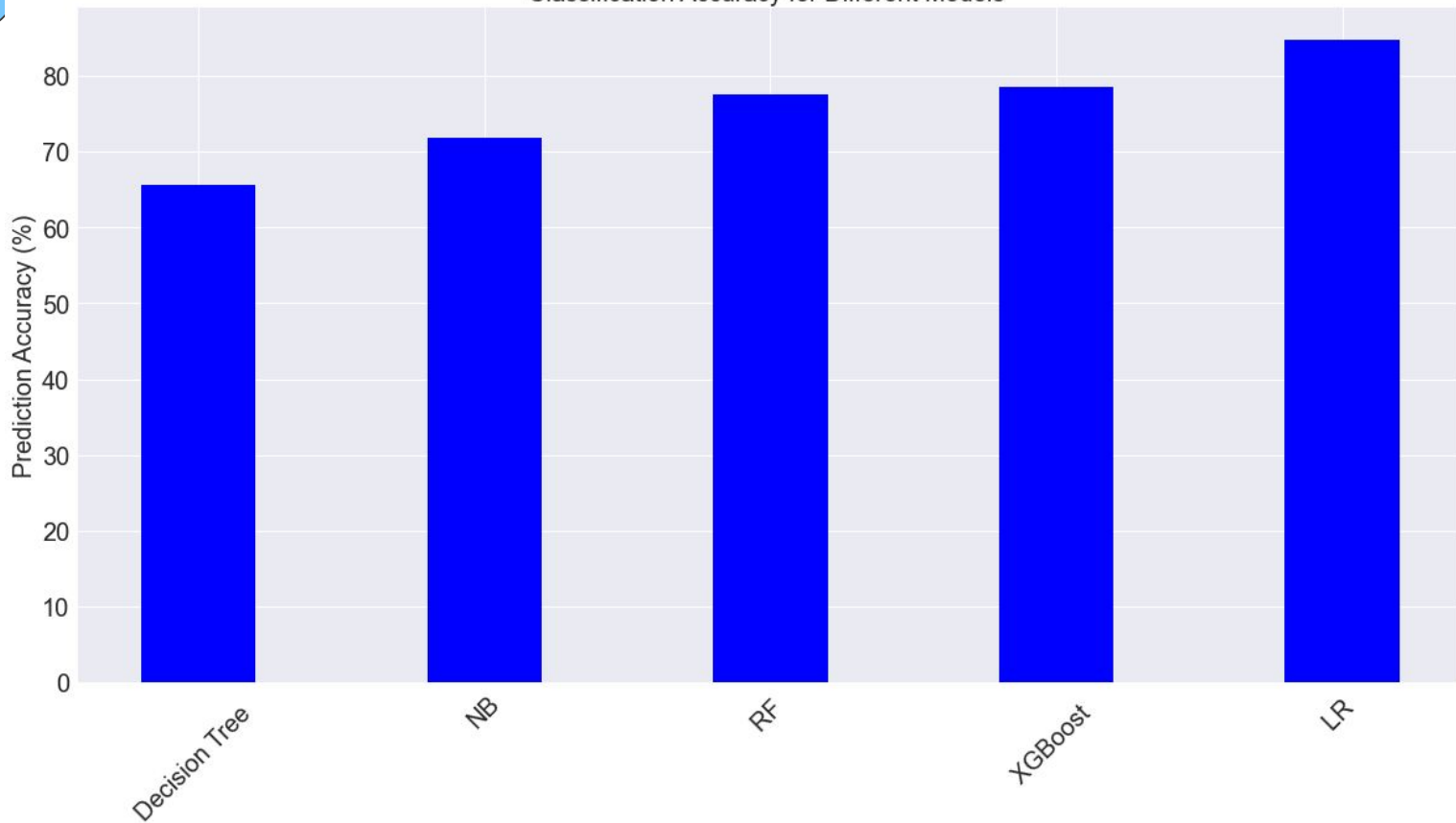


Prediction Accuracy

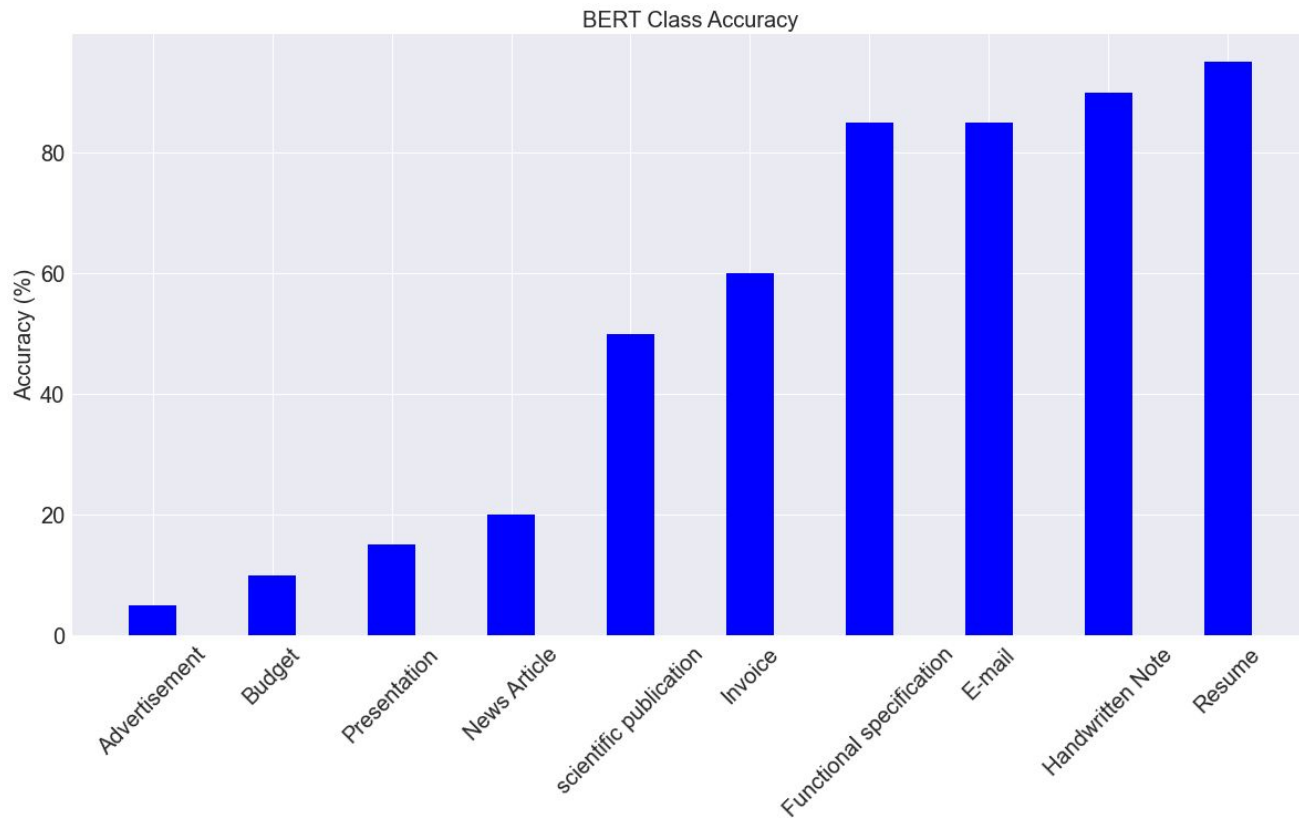
In-sample Dev test: 84%
Out-of-sample test: 90%

Model Evaluation

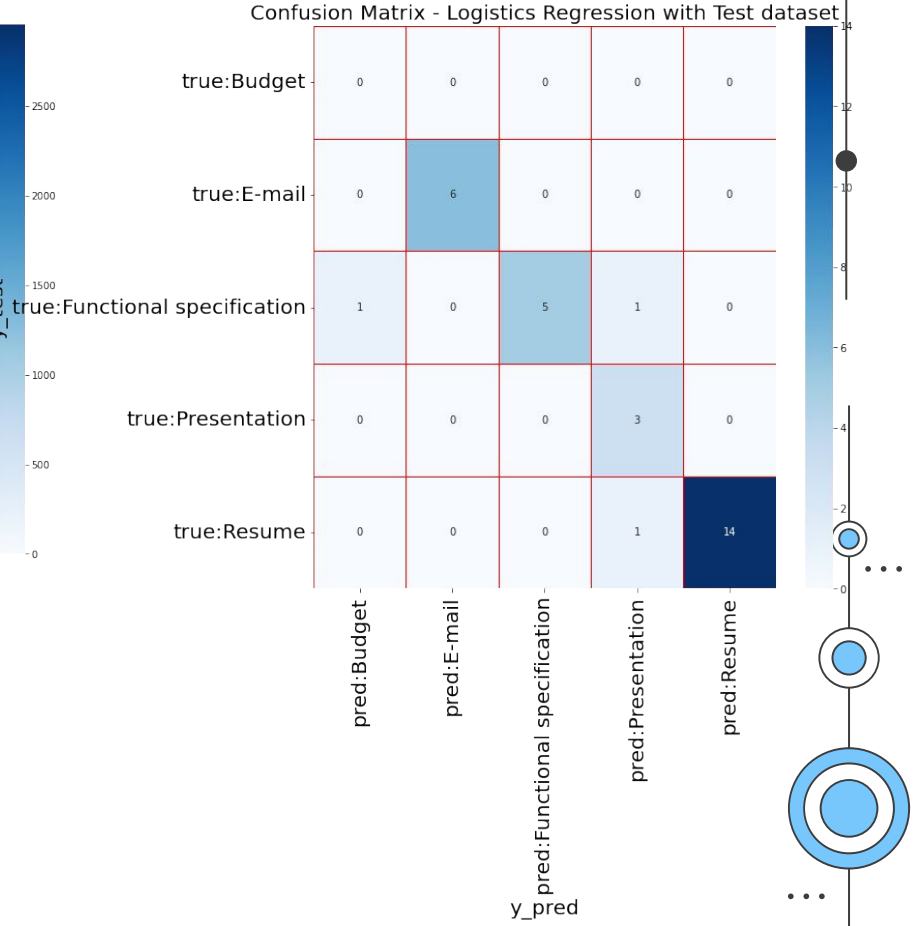
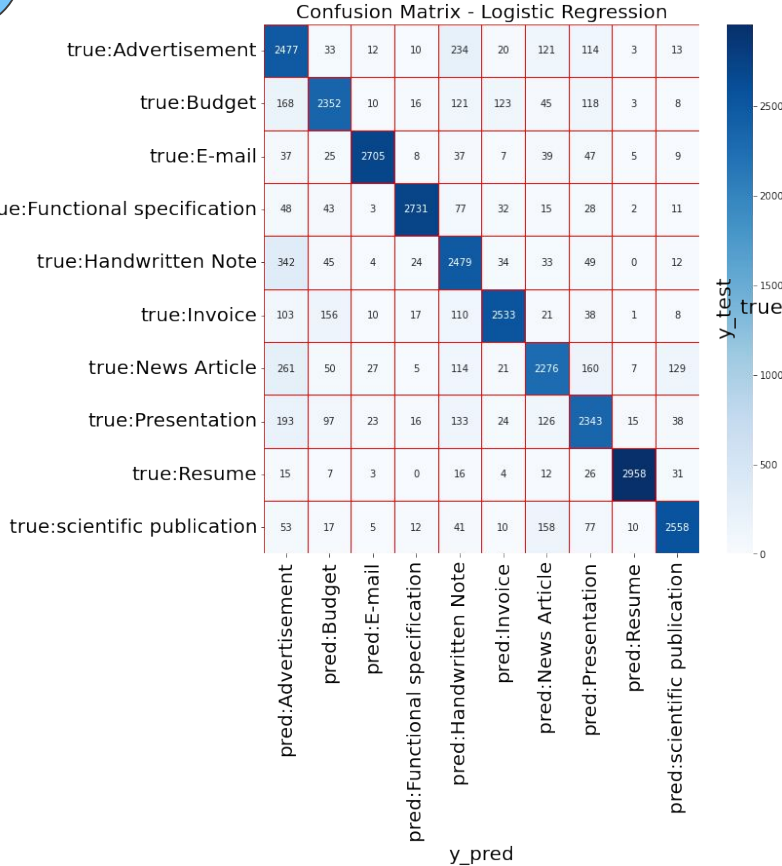
Classification Accuracy for Different Models



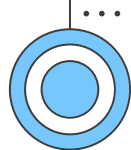
BERT Evaluation (Small Data Set)



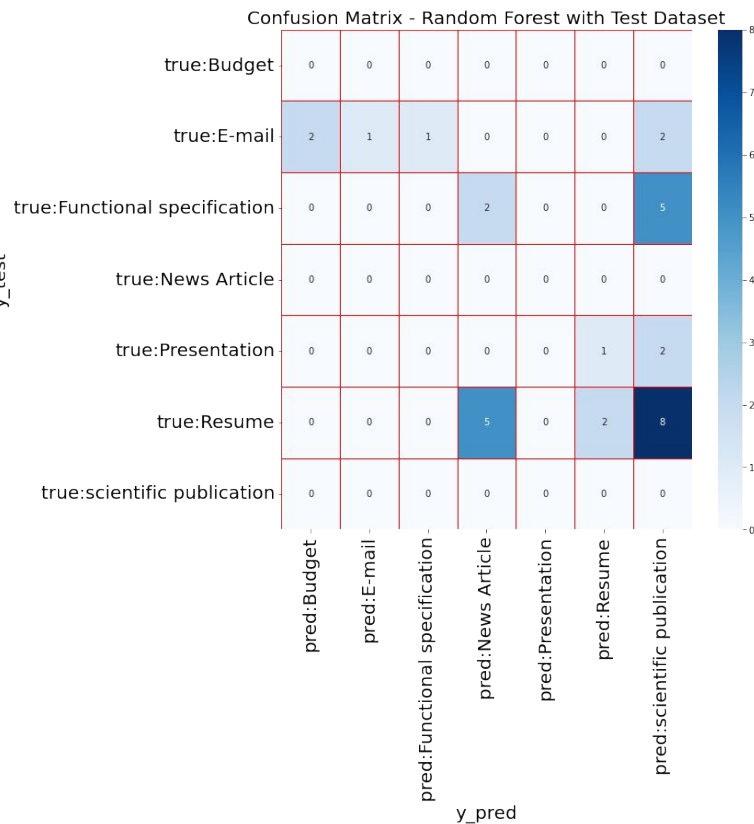
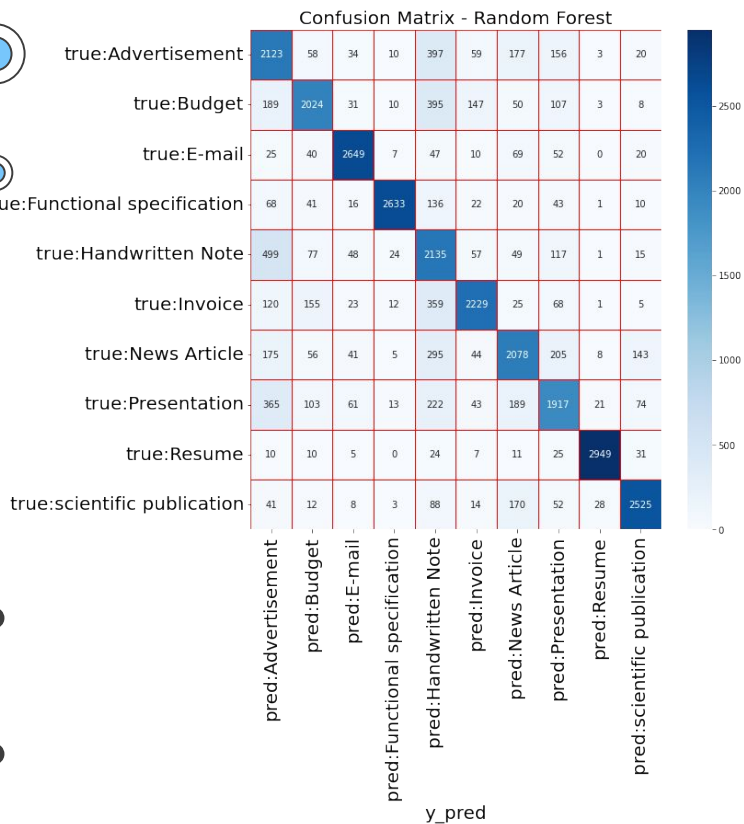
Model Evaluation (LR)



Module Evaluation (RF)



...
y_test



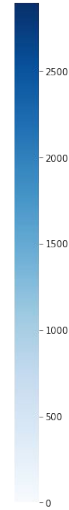
Module Evaluation(NB)

Confusion Matrix - Naive Bayes

	pred:Advertisement	pred:Budget	pred:E-mail	pred:Functional specification	pred:Handwritten Note	pred:Invoice	pred:News Article	pred:Presentation	pred:Resume	pred:scientific publication
true:Advertisement	1010	56	67	62	151	44	1460	75	17	95
true:Budget	4	1956	67	26	115	104	502	79	33	78
true:E-mail	1	23	2527	14	6	2	235	60	8	43
true:Functional specification	1	13	18	2650	39	8	186	6	2	67
true:Handwritten Note	20	35	101	45	1526	48	1011	11	11	214
true:Invoice	0	229	30	20	66	2204	376	20	10	42
true:News Article	9	16	10	7	31	7	2761	56	10	143
true:Presentation	8	178	56	39	124	30	745	1488	56	284
true:Resume	0	0	2	0	2	2	79	57	2897	33
true:scientific publication	1	3	5	1	8	4	281	7	86	2545

y_test

y_pred



Confusion Matrix - Naive Bayes with test dataset

	pred:E-mail	pred:Functional specification	pred:News Article	pred:Presentation	pred:Resume
true:E-mail	5	0	1	0	0
true:Functional specification	2	1	0	4	0
true:News Article	0	0	0	0	0
true:Presentation	0	0	0	3	0
true:Resume	0	0	2	9	4

y_test

y_pred

