





## Achieving Harmony in Code Declarations through Web Scraping and ML

---

**ACRA**  
redefined.

- 
- The diagram for Phase 1 features a large, hollow black outline of the number '1' on the left. To its right is a solid orange circle. A thick black line forms a semi-circle above and below the circle, connecting to the '1'. Inside the orange circle is a bulleted list of six items. At the bottom center of the circle are three small white dots.
- Business Understanding
  - Project Planning and Charter
  - Data Source and EDA
  - Data Extraction and Pre-processing
  - Initial Modelling
  - Retrospective and Plans for Phase II

- 
- The diagram for Phase 2 features a large, hollow black outline of the number '2' on the left. To its right is a solid blue circle. A thick black line forms a semi-circle above and below the circle, connecting to the '2'. Inside the blue circle is a bulleted list of six items. At the bottom center of the circle are three small white dots.
- LLM Exploration
  - Model Evaluation and Comparison
  - Iterative Refinement
  - Project Delivery
  - Retrospective and Project Conclusion

# Objective



From SSIC Recommender Project

Current State

Entity	Declared SSIC Code
Entity_A	SSIC_001

- Entity\_A declared SSIC\_001
- No validation of SSIC Code and Entity\_A's Business Description

Target End State

Entity	Declared SSIC Code	Recommended SSIC Codes	Validation Score
Entity_A	SSIC_001	SSIC_001, SSIC_002	0.85*

- Entity\_A declared SSIC\_001
- Get list of Recommended/Classified SSIC Codes based on Business Description of Entity\_A
- Validation of Declared and Recommended SSIC Codes
- Validation Score to measure SSIC misalignment, allow quick filter and identification for further investigation

# Objective

## Methodology

1

Explore Sources of Entity's Business Description

- Availability/Accessibility/Extractability
- Relevance

2

Explore Types of Model/Recommendation/Classification Techniques

- **Recommendation System (RCS)**
- Large Language Learning Model (LLM)
  - Natural Language Inference (NLI), PST Ideasfest
  - Prompt Engineering, GoogleTrailblazer
- **Text Classification**

a

Dataset/Reference/Dictionary/Corpus



Model

### Current State

Entity	Declared SSIC Code	Entity's Business Description
Entity_A	SSIC_001	Entity_A's Business Description

- Entity\_A declared SSIC\_001
- No validation of SSIC Code and Entity\_A's Business Description

### Target End State

Entity	Declared SSIC Code	Recommended SSIC Codes	Validation Score
Entity_A	SSIC_001	SSIC_001, SSIC_002	0.85*

- Entity\_A declared SSIC\_001
- Get list of Recommended/Classified SSIC Codes based on Business Description of Entity\_A
- Validation of Declared and Recommended SSIC Codes
- Validation Score to measure SSIC Misalignment, allow quick filter and identification for further investigation

# Summary

## Data Source & Understanding

Source	DOS	ACRA	Online	
<i>Source Data</i>	- Industries by Alphabetical Index - Detailed Definitions	Annual Reports	LinkedIn Organization Page	Organization Main/Sub Pages
<i>Accessibility (and Availability)</i>	★★★★ Singstat (DOS) SSIC Website	★★★★ ACRA's DB	★ - Not every Organization have - Extracted via Web scraping - Security Bot (Captcha)	★ - Not every Organization have - Extracted via Web scraping
<i>Relevancy</i>	★★★★ - Maintained by DOS - SSIC Definition updated every 5 years	★★★★ Updated annually*	★★ - Assume page is updated	★★ - Assume page is updated
<i>Overall</i>	★★★★	★★★★	★	★★

Good
Subjective
Non-ideal

### Recommendations

1. Annual Reports **main** source of Business Activity Data
2. Organization & LinkedIn Webpages **alternative** sources

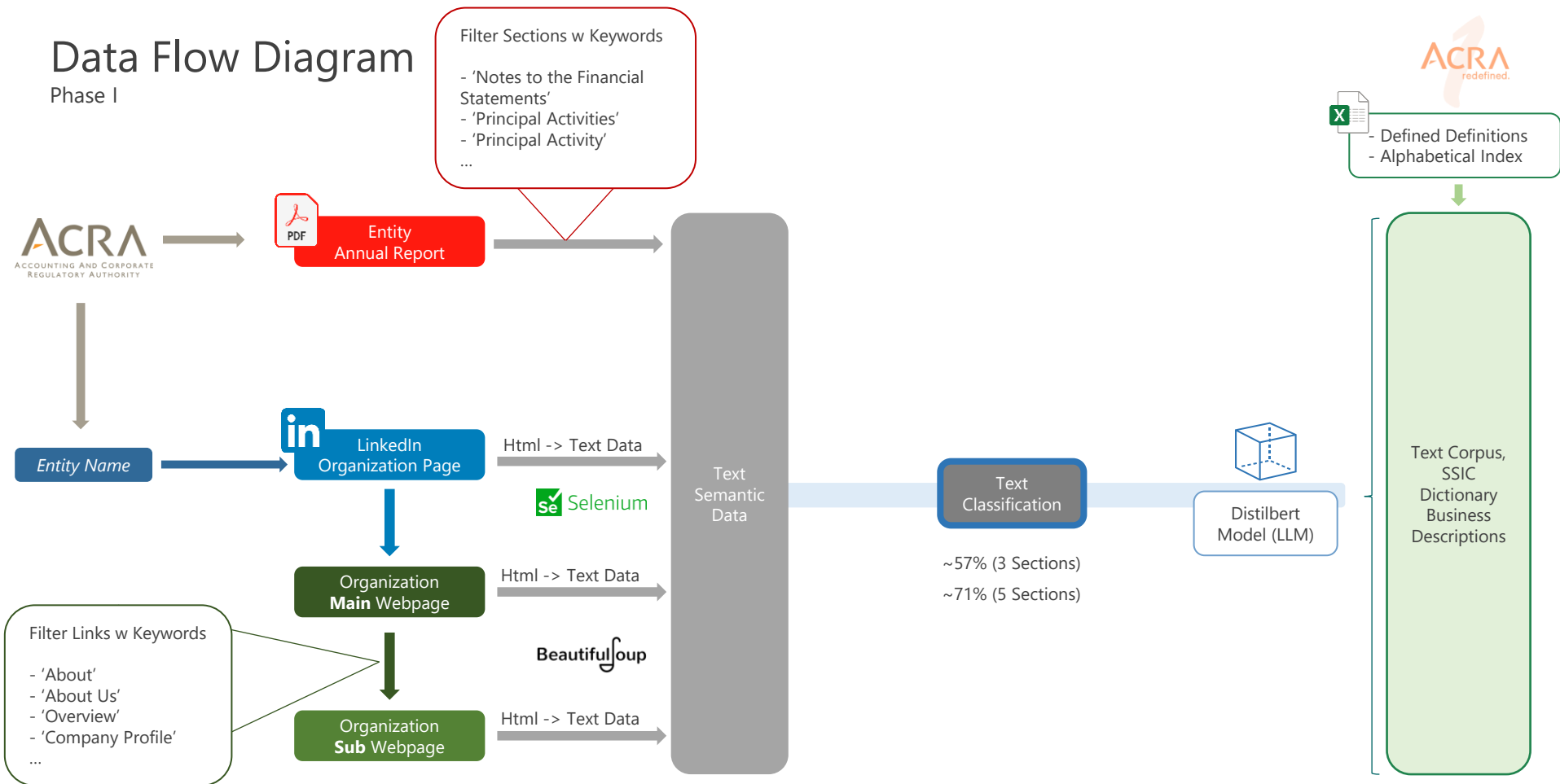
# Summary

## Model Types

Model	Recommender System	Large Language Model		
<i>Sub Model Type</i>	<b>Content-based</b>	<b>Natural Language Inference</b>	<b>Prompt Engineering</b>	<b>DistilBert</b>
<i>Pros</i>	<ul style="list-style-type: none"><li>Personalized recommendations based on user preferences</li><li>Easy to understand and explain</li><li>No cold-start problem for new items.</li></ul>	<ul style="list-style-type: none"><li>Enhances understanding of complex relationships between texts</li><li>Supports various downstream NLP tasks</li><li>Improves question-answering systems.</li></ul>	<ul style="list-style-type: none"><li>Customizes language model behavior</li><li>Enhances performance on specific tasks</li><li>Low-cost adjustment without retraining.</li></ul>	<ul style="list-style-type: none"><li>Faster and more efficient than BERT</li><li>Good performance with less computational resources</li><li>Suitable for deployment in resource-constrained environments.</li></ul>
<i>Cons</i>	<ul style="list-style-type: none"><li>Limited to user's existing preferences</li><li>Cannot recommend novel items outside user's profile</li><li>Requires extensive feature engineering.</li></ul>	<ul style="list-style-type: none"><li>Requires large annotated datasets</li><li>Computationally intensive</li><li>May struggle with nuanced or ambiguous language.</li></ul>	<ul style="list-style-type: none"><li>Requires expertise to design effective prompts</li><li>Limited by the language model's original training</li><li>Trial and error process can be time-consuming.</li></ul>	<ul style="list-style-type: none"><li>Slightly lower accuracy compared to full BERT</li><li>May require fine-tuning for specific tasks</li></ul>
<i>Overall</i>				

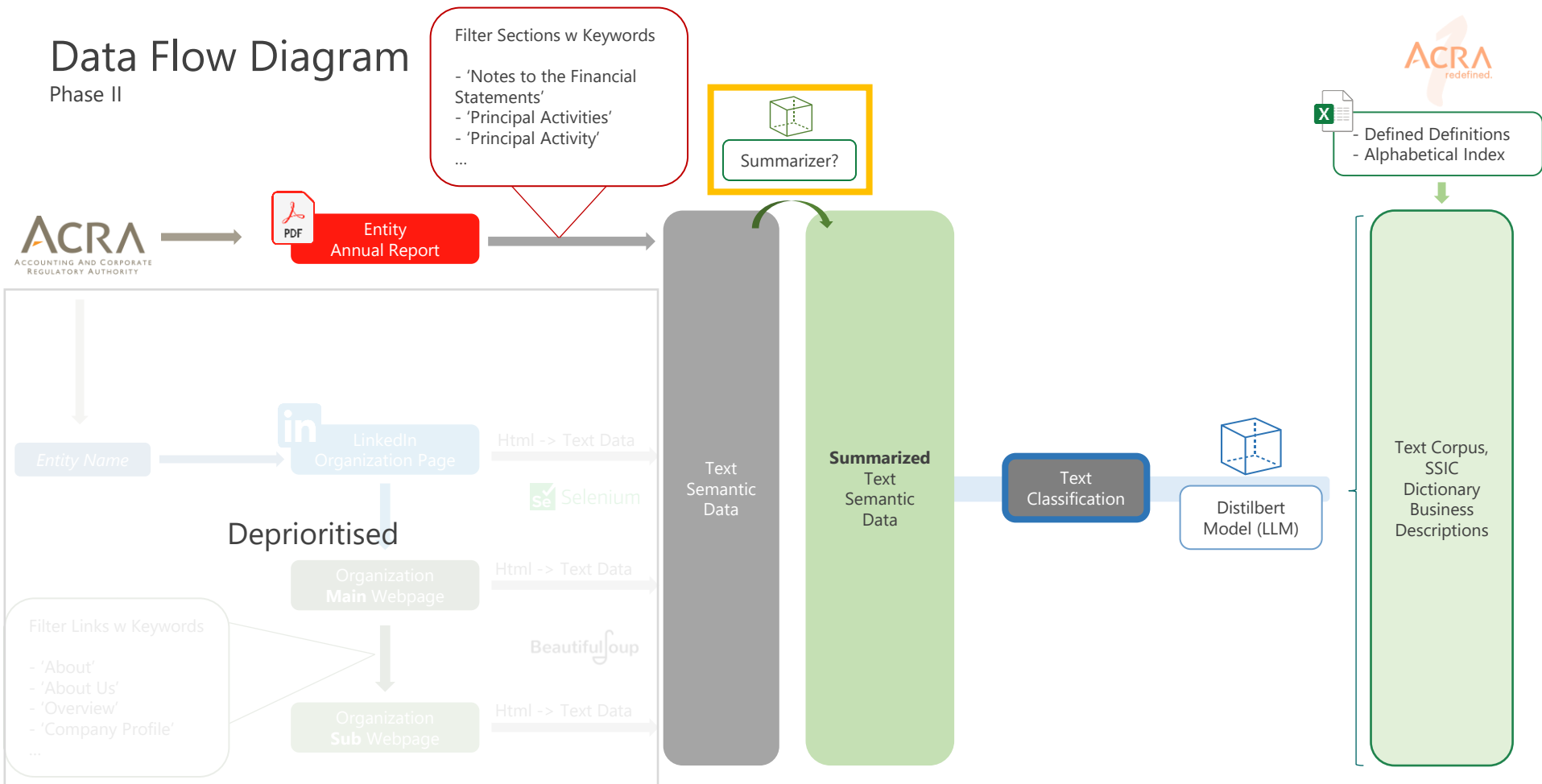
# Data Flow Diagram

Phase I



# Data Flow Diagram

Phase II





# Google Trailblazer

Digital Assistant for SSIC Deviation Detection



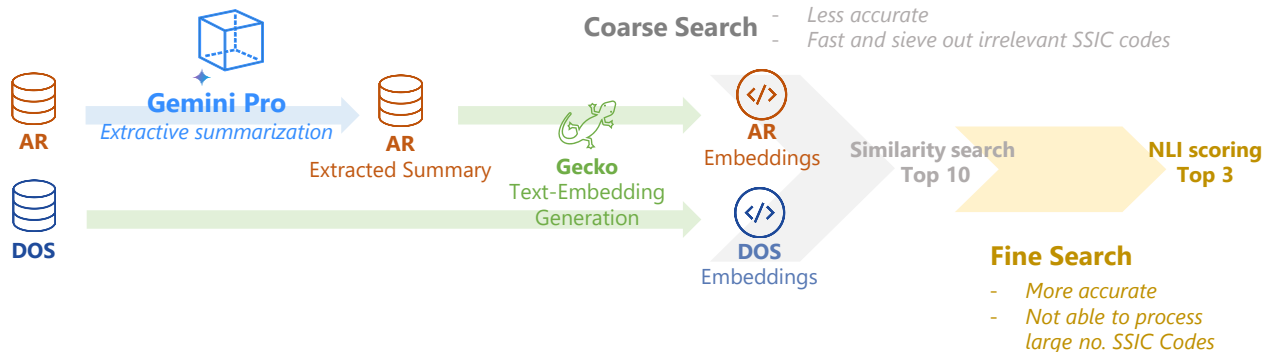
Dictionary list of  
SSIC and Descriptions  
(Validated)



Annual Report PDF  
(Unvalidated)

## Model

Prompt Engineering



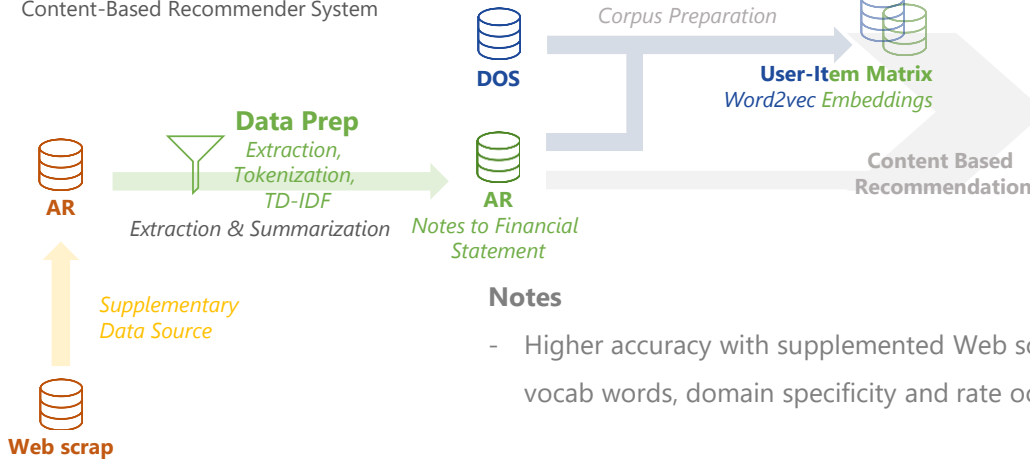
# Phase I

## Initial Exploration and Model



### 1<sup>st</sup> Model

Content-Based Recommender System



#### Validation @ Section Level

\*AR (wo WS): 36%  
\*AR (w WS): **56%**

\*Assumption AR is validated

#### Notes

- Higher accuracy with supplemented Web scrap data, due wider range of vocab words, domain specificity and rate occurrences

#### Limitations

- Limited to predefined features with feature engineering
- Webscrap info cleanup effort, integrity and relevance
- Domain specific and struggle with ambiguous/out-of-context terms
- Heavy reliance on exact matches or predefined synonyms
- **Underlying assumption of validated AR**

# Phase II

## LLM Exploration



DOS

Dictionary list of  
SSIC and Descriptions  
(Validated)



AR

Annual Report PDF  
(Unvalidated)



Web scrap

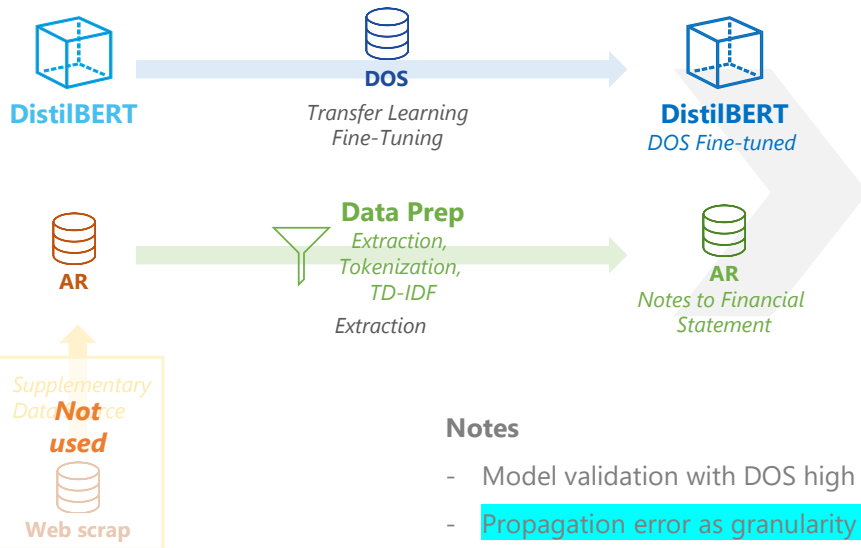
Web scrap LinkedIn &  
Corporate Website  
(Unvalidated)

**Not  
used**

### 2<sup>nd</sup> Model

Fine-Tuning DistilBERT for Multiclass Text Classification

5 Different Models for each Level



Validation w <b>Section</b> model			
Section,	DOS: <b>99%</b>	*AR: <b>69%</b>	
Validation w <b>Group</b> model			
Section,	DOS: <b>99%</b>	*AR: <b>67%</b>	
Division,	DOS: <b>98%</b>	*AR: 50%	
Group,	DOS: <b>96%</b>	*AR: <b>24%</b>	

\*Assumption AR is validated

### Notes

- Model validation with DOS high and consistent
- Propagation error as granularity increase (Section vs Group model)
- Different semantic input between DOS and AR Notes to the Financial Statement
- Observed noise/irrelevant data from AR Notes to the Financial Statement

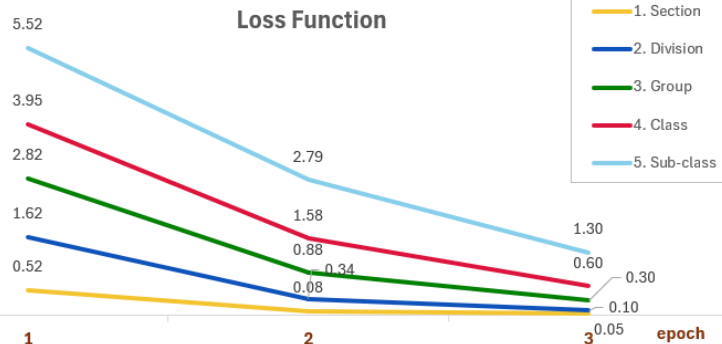
# DistilBERT

5 model for each level

Hyperparameter	
Optimizer	Adam
Learning Rate	5.00E-05
Epsilon	1.00E-08
Epoch	3
Batch Size	16

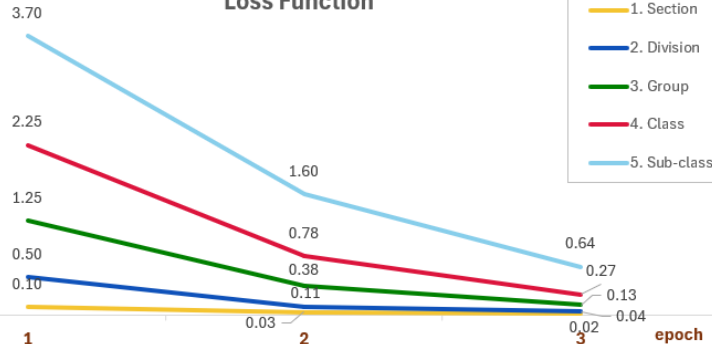
Training Dataset

Loss Function

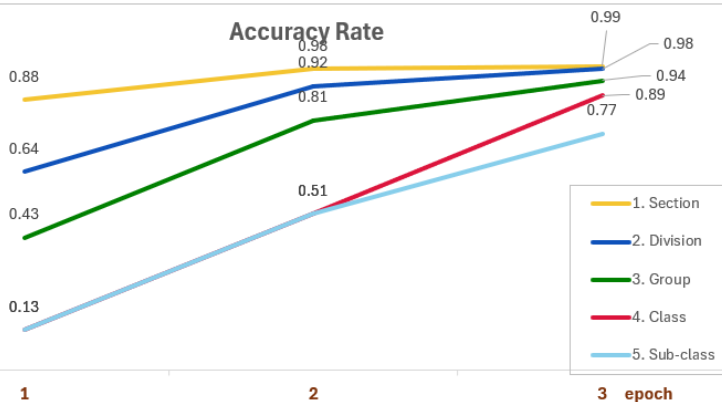


Testing Dataset

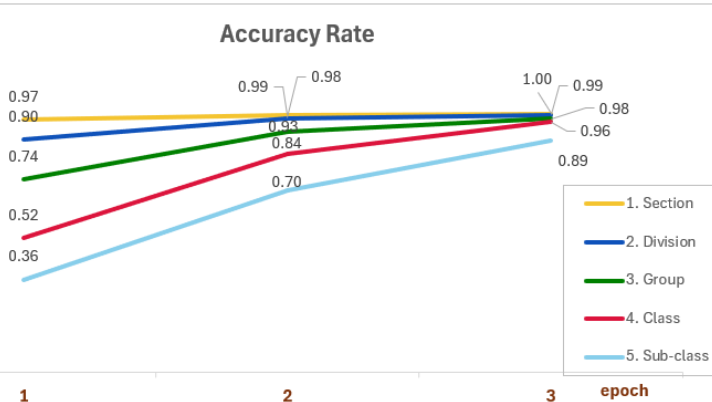
Loss Function



Accuracy Rate



Accuracy Rate



# Phase II

## LLM Exploration



DOS

Dictionary list of  
SSIC and Descriptions  
(Validated)



AR

Annual Report PDF  
(Unvalidated)



Web scrap

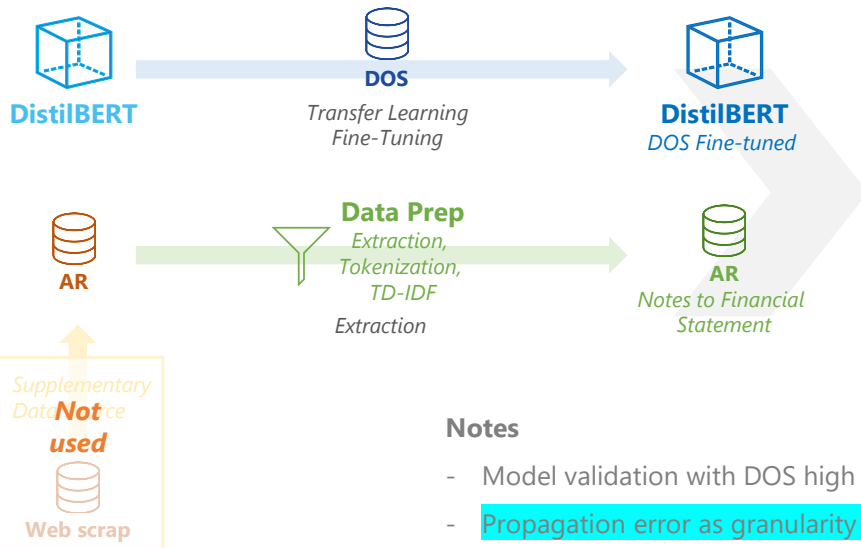
Web scrap LinkedIn &  
Corporate Website  
(Unvalidated)

**Not  
used**

### 2<sup>nd</sup> Model

Fine-Tuning DistilBERT for Multiclass Text Classification

5 Different Models for each Level



Validation w <b>Section</b> model			
Section,	DOS: <b>99%</b>	*AR: <b>69%</b>	
Validation w <b>Group</b> model			
Section,	DOS: <b>99%</b>	*AR: <b>67%</b>	
Division,	DOS: <b>98%</b>	*AR: 50%	
Group,	DOS: <b>96%</b>	*AR: <b>24%</b>	

\*Assumption AR is validated

### Notes

- Model validation with DOS high and consistent
- Propagation error as granularity increase (Section vs Group model)
- Different semantic input between DOS and AR Notes to the Financial Statement
- Observed noise/irrelevant data from AR Notes to the Financial Statement

# Phase II

## LLM Exploration & Summarization



Dictionary list of  
SSIC and Descriptions  
(Validated)



Annual Report PDF  
(Unvalidated)

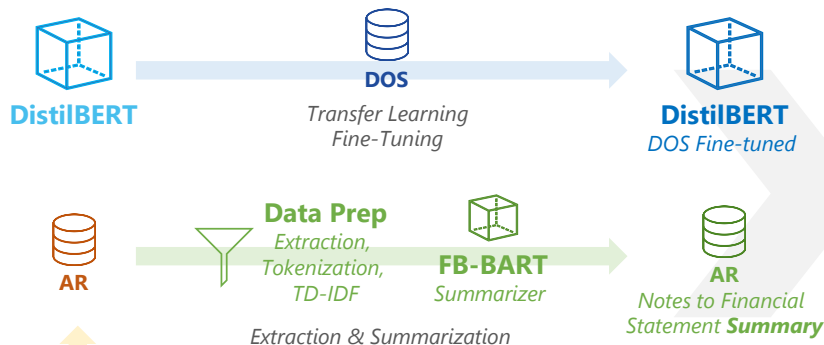


Web scrap  
**Not used**  
LinkedIn &  
Corporate Website  
(Unvalidated)

### 2<sup>nd</sup> Model

Fine-Tuning DistilBERT for Multiclass Text Classification

5 Different Models for each Level



Validation w <b>Section</b> model			
Section,	DOS: <b>99%</b>	*AR: <b>74%</b>	
Validation w <b>Group</b> model			
Section,	DOS: <b>99%</b>	*AR: <b>70%</b>	
Division,	DOS: <b>98%</b>	*AR: <b>61%</b>	
Group,	DOS: <b>96%</b>	*AR: <b>40%</b>	

\*Assumption AR is validated

### Notes

- FB-BART pre-trained on news articles and Wikipedia, formal structured texts

### Limitations

- Model training/fine-tuning (CPU) ~2-3hr (subject to hyperparam and level)
- Underlying assumption of validated AR

# Github

## ssicsync code repo



README

## Background

The Singapore Standard Industrial Classification (SSIC) is a critical numerical 5-digit coding system used to classify economic activities in Singapore. It serves as a key indicator in various surveys and databases, providing insights into the economic landscape. When businesses commence operations in Singapore, they are mandated to declare an appropriate SSIC code. However, due to the vast diversity of SSIC codes and the complexity of business activities, ensuring the accuracy of these declared SSIC codes presents a significant challenge.

The primary aim of this project is to enhance the accuracy of SSIC code verification and declarations. We seek to achieve this by using state of the art Natural Language Processing (NLP) techniques.

## How to use

1. Git clone this repository and pull from 'main' branch.
2. Update [list of companies](#) to predict SSIC codes.
3. Upload companies' annual reports [here](#). Insert company's UEN number in the file names [e.g., ABC PTE LIMITED (199999999C).pdf].
4. Ensure that SSIC Code's [reference files](#) are up to [date](#).
5. Create and activate virtual environment ("python -m venv myenv" followed by "myenv\Scripts\activate").
6. Install required packages ("pip install -r requirements\_repo.txt").
7. If you wish to train the transfer learning models, run "python training.py" and upload model files to [Hugging Face](#). Change the hard-coded values at the top of script to your preference.
8. Run "python main.py" to generate predicted SSIC results for the list of companies. Change the hard-coded values at the top of script to your preference.
9. Push updated files back to main branch (*important*). Exclude model files.
10. Visualize results on [Streamlit](#).

ssicsync Public

Pin Unwatch 1 Fork 0 Star 0

main 8 Branches 0 Tags

Go to file Add file Code

Michael Wong	update	09a7d62 · 2 weeks ago	461 Commits
dataSources	update	2 weeks ago	
images	update repo	2 weeks ago	
logs	update	2 weeks ago	
models	update	2 weeks ago	
pages	Update 1_Results.py	2 weeks ago	
results	update	2 weeks ago	
sandbox	update	2 weeks ago	
.gitignore	Initial Commit	3 months ago	
README.md	update	2 weeks ago	
commonFunctions.py	Create distilbert model training copy.ipynb	3 weeks ago	
controller.py	update - to test end to end!	2 weeks ago	
main.py	update	2 weeks ago	
requirements.txt	update repo	2 weeks ago	
requirements_repo.txt	update repo	2 weeks ago	
training.py	update hard-coded values	2 weeks ago	
_Homepage.py	update repo	2 weeks ago	

### About

No description, website, or topics provided.

Readme Activity 0 stars 1 watching 0 forks

### Releases

No releases published  
[Create a new release](#)

### Packages

No packages published  
[Publish your first package](#)

### Contributors

### Languages

Jupyter Notebook 98.5%

Python 1.5%

### Suggested workflows

Based on your tech stack

<https://github.com/yorwel/ssicsync>



**nusebacra**

nusebacra

Follow



AI & ML interests

None yet

Organizations

None yet

Models 5



Sort: Recently updated

nusebacra/ssicsync\_subclass\_classifier

Text Classification • Updated Jun 25 • ↓ 97

nusebacra/ssicsync\_class\_classifier

Text Classification • Updated Jun 24 • ↓ 34

nusebacra/ssicsync\_group\_classifier

Text Classification • Updated Jun 24 • ↓ 85

nusebacra/ssicsync\_division\_classifier

Text Classification • Updated Jun 24 • ↓ 35

nusebacra/ssicsync\_section\_classifier

Text Classification • Updated Jun 24 • ↓ 45


Datasets

None public yet

<https://huggingface.co/nusebacra>



[Homepage](#)
[Results](#)
[Prediction \(Section\)](#)
[Prediction \(Division\)](#)
[Prediction \(Group\)](#)
[Prediction \(Class\)](#)
[Prediction \(Sub-class\)](#)
[Reference \(Section\)](#)
[Reference \(Division\)](#)
[Reference \(Group\)](#)
[Reference \(Class\)](#)
[Reference \(Sub-class\)](#)



## About this Webpage

This platform offers an interactive exploration of SSIC classification results, from overall accuracy metrics to detailed company-level analyses. Users can leverage the Prediction pages to input custom company descriptions, allowing the model to generate and return the most relevant SSIC codes based on the specified hierarchical level. Additionally, the Reference pages provide a quick search feature for SSIC codes, enabling users to gain a deeper understanding of their applications.

## Table of Contents

**Results**

This section presents the overall classification results as well as SSIC results at the company level. It is particularly useful for validating companies' declared SSIC codes against the recommended SSIC codes.

**Prediction (Section)**

This section enables users to apply the classification model to ad-hoc company descriptions, returning the top SSIC codes at the Section level. It is ideal for conducting quick analyses to obtain the recommended SSIC codes at the Section level.

**Prediction (Division)**

This section enables users to apply the classification model to ad-hoc company descriptions, returning the top SSIC codes at the Division level. It is ideal for conducting quick analyses to obtain the recommended SSIC codes at the Division level.

**Prediction (Group)**

This section enables users to apply the classification model to ad-hoc company descriptions, returning the top SSIC codes at the Group level. It is ideal for conducting quick analyses to obtain the recommended SSIC codes at the Group level.

**Prediction (Class)**

This section enables users to apply the classification model to ad-hoc company descriptions, returning the top SSIC codes at the Class level. It is ideal for conducting quick analyses to obtain the recommended SSIC codes at the Class level.

Level of Classification:

Section

	Company Name	Adjusted Score	Within Top 3
0	ABR HOLDINGS LIMITED	0.28	Yes
1	ABUNDANCE INTERNATIONAL LIMITED	0.28	Yes
2	ABUNDANTE LIMITED	0.28	Yes
3	ACCRELIST MEDICAL AESTHETICS (SPC) PTE. LTD	0.00	No
4	ACESIAN PARTNERS LIMITED	0.13	Yes
6	ADVANCED SYSTEMS AUTOMATION LIMITED	0.00	No
7	ALLIANCE HEALTHCARE GROUP LIMITED	0.00	No
8	ANNAIK LIMITED	0.00	No
9	AP OIL INTERNATIONAL LIMITED	0.00	No
10	BEST WORLD INTERNATIONAL LIMITED	0.13	Yes

List of Companies

ABR HOLDINGS LIMITED

## Company SSIC Details

**Company Name:**

ABR HOLDINGS LIMITED

**Company Adjusted Score:**

0.28

**Company Description:**

Principal activities of the company are the manufacture of ice cream, the operation of swensens ice cream parlours cum restaurants, operation of other specialty restaurants and investment holding.

**Company SSICs & Descriptions:**

G: Wholesale and retail trade  
K: Financial and Insurance activities

**Top 3 Predicted SSICs & Descriptions:**

I: Accommodation and food service activities  
C: Manufacturing  
G: Wholesale and retail trade

ABR HOLDINGS LIMITED SSIC Codes are within its predicted top 3 SSIC Codes.

\*The Adjusted Score is a metric designed to assign higher weights to a company's top SSIC predictions, with progressively lower weights applied to subsequent predictions. Weights are also distributed in descending order from Sub-class to Class, Group, Division, and Section. These weights are then aggregated to compute the overall Adjusted Score for each company. This score measures accuracy based on the top predictions, regardless of classification level. The Adjusted Score ranges from 0 to 1, where a value closer to 0 indicates poorer overall classification accuracy, and a value closer to 1 indicates stronger overall classification accuracy.

<https://ssicsync-nwdmvmh4vzhx4yfqphazs.streamlit.app/>

## Business Description Classifier

### Classification (1032 Sub-class Categories)

Welcome to the Business Description Classifier! This application utilizes a multi-class text classification model to categorize business descriptions into one of 1032 Sub-class categories. Simply input your business description, and the model will analyze the text and provide a list of predicted categories.

### How to Use

1. Enter the business description in the text box below.
2. Hit Control + Enter.
3. The top 3 predicted categories will be displayed below the button.

Enter Business Description:

### About the Model

This model has been trained on a diverse dataset of business descriptions and is capable of understanding and classifying a wide range of business activities. The 1032 Sub-class categories cover various industry sectors, providing accurate and meaningful classifications for your business needs.

### Examples

- **Technology:** Software development, IT consulting, hardware manufacturing.
- **Healthcare:** Hospitals, pharmaceutical companies, medical research.
- **Finance:** Banking, insurance, investment services.

## Sub-class, 1032 Categories

Search by Sub-class:

Search by Title Keywords:

### Sub-class Reference Table:

	Sub-class	Sub-class Title
0	01111	Growing of leafy and fruit vegetables
1	01112	Growing of mushrooms
2	01113	Growing of root crops
3	01119	Growing of food crops (non-hydroponics) n.e.c.
4	01120	Growing of leafy and fruit vegetables (hydroponics)
5	01130	Growing of fruits
6	01141	Growing of orchids
7	01142	Growing of ornamental plants
8	01149	Growing of nursery products n.e.c.
9	01190	Growing of other crops

<https://ssicsync-nwdmvmh4vzhx4yfqphazs.streamlit.app/>



# Retrospective

*"We do not learn from experience...  
we learn from reflecting on experience."*

*– John Dewey*

# Challenges I

## Data Sourcing and Extraction

### Web scraping in Phase I

- Digital/Online presence not observed in all companies
- Ethical concerns of web scraping by government bodies
- Data source shift to more reliable source **Annual Report\***



No LinkedIn Profile or Corporate Website

### "Notes to the Financial Statement"

- Other complementing sections (e.g. subsidiaries)
- Irregular layouts, diverse Formatting (e.g. tables)
- Alternatives sources** (Financial Statements & XBRL)

NOTES TO THE FINANCIAL STATEMENTS	
31 March 2023	
1. GENERAL	
S&S Engineering Company Limited (the "Company") is a limited liability company incorporated in the Republic of Singapore which is also the place of domicile. The Company is listed on the Singapore Exchange Securities Trading Limited ("SGX-ST"). The Company is a subsidiary company of Singapore Airlines Limited and its ultimate holding company is Temasek Holdings (Private) Limited. Both holding companies are incorporated in the Republic of Singapore.	
The registered office of the Company is at 31 Ardmore Road, Singapore 618811.	
The financial statements of the Group as at 31 March 2023 and for the year then ended comprise the Company and its subsidiary companies (together referred to as the "Group" and individually as "Group entities") and the Group's interest in equity accounted investees.	
The principal activities of the Company are the provision of aviation maintenance, component overhaul services and aircraft technical management, the provision of the maintenance and technical ground handling services and passenger handling. The principal activities of the subsidiary companies are described in Notes 16 to the financial statements. There have been no significant changes in the nature of these activities during the financial year.	
The financial statements for the financial year ended 31 March 2023 were audited for issue in accordance with a resolution of the Board of Directors on 8 May 2023.	
2. SUMMARY OF SIGNIFICANT ACCOUNTING POLICIES	

Notes to the Financial Statement

The associated companies are:			
	Principal activities	Country of incorporation and place of business	Percentage of equity held by the Group 31 March 2023 2022
Held by the Company			
TATA SIA Airlines Limited <sup>(a)</sup>	Domestic and international full service scheduled passenger airlines services	India	49.0 49.0
Airbus Asia Training Centre Pte. Ltd. <sup>(a)</sup>	Flight training services	Singapore	45.0 45.0
Ritz-Carlton, Millenia Singapore Properties Private Limited <sup>(a)</sup>	Hotel ownership and management	Singapore	20.0 20.0
Held by SIAEC			
Boeing Asia Pacific Aviation Services Pte. Ltd. <sup>(a)</sup>	Provide engineering, material management and fleet support solutions	Singapore	38.0 38.0
Eagle Services Asia Private Limited <sup>(a)</sup>	Repair and overhaul of aircraft engines	Singapore	38.0 38.0
Fuel Accessory Service Technologies Pte Ltd <sup>(a)</sup>	Repair and overhaul of engine fuel components and accessories	Singapore	38.0 38.0
GE Aviation, Overhaul Services - Singapore Pte. Ltd. <sup>(a)</sup>	Repair and servicing of aircraft and spacecraft (including aircraft engines and other parts)	Singapore	38.0 38.0
Moog Aircraft Services Asia Pte. Ltd. <sup>(a)</sup>	Repair and overhaul services for flight control systems	Singapore	38.0 38.0
PT JAS Aero-Engineering Services <sup>(a)</sup>	Provide aircraft maintenance services, including technical and non-technical handling at the airport	Indonesia	38.0 38.0
Southern Airports Aircraft Maintenance Services Company Limited <sup>(a)</sup>	Provide aircraft maintenance services, including technical and non-technical handling at the airport	Vietnam	38.0 38.0
Component Aerospace Singapore Pte. Ltd. <sup>(a)</sup>	Repair and overhaul of aircraft engine combustion chambers, guides, fuel nozzles and related parts	Singapore	36.0 36.0
JAMCO Aero Design & Engineering Private Limited <sup>(a)</sup>	Providing turnkey solutions for aircraft interior modifications	Singapore	34.9 34.9
Panasonic Avionic Services Singapore Pte. Ltd. <sup>(a)</sup>	Provide line maintenance and repair services of in-flight entertainment systems	Singapore	32.9 32.9
Goodrich Aerostructures Service Centre-Asia Pte. Ltd. <sup>(a)</sup>	Repair and overhaul of aircraft nacelles, thrust reverses and pylons	Singapore	31.0 31.0
Pan Asia Pacific Aviation Services Limited <sup>(a)</sup>	Provide aircraft maintenance services, including technical and non-technical handling at the airport	Hong Kong	31.0 31.0

Subsidiary Companies (Holding Companies)

# Challenges II

## Nature of SSIC Codes

### - Unvalidated

- Self-declared and not assigned
- No true SSOT or validation process
- Lack of validated annotated data

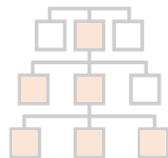


### - Transient

- Temporary and subject to change
- Entity's autonomy for declaration and changes
- Assumption of Highest revenue = Primary SSIC

## Multitude of Classification Categories

- Up to 2 SSIC per Entity (*Primary & Secondary*)
- Up to 1,000 at Sub-class level
- Risk of Error Propagation
- **Group** level classification



### Structure of the Classification

2.7. The SSIC is a classification with a hierarchical structure. At the highest level of aggregation, there are 21 broad categories known as "Sections", each denoted by a single alphabetical letter. Each "Section" comprises one or more "Divisions" as shown below:

Section	Division
A Agriculture and Fishing	01-03
B Mining and Quarrying	08-09
C Manufacturing	10-32
D Electricity, Gas, Steam and Air-Conditioning Supply	35
E Water Supply; Sewerage, Waste Management and Remediation Activities	36-38
F Construction	41-43
G Wholesale and Retail Trade	46-47
H Transportation and Storage	49-53
I Accommodation and Food Service Activities	55-56
J Information and Communications	58-63
K Financial and Insurance Activities	64-66

### Sub-class 01130 'Growing of fruits'

This Sub-class includes growing of fruits such as banana, papaya, mangoes, dates and/or pineapples.

This Sub-class excludes:

- growing of leafy and fruit vegetables (non-hydroponics), see 01111
- growing of leafy and fruit vegetables (hydroponics), see 01120
- growing of nursery products (e.g. orchids, ornamental plants), see 0114

### Section A: Agriculture and Fishing

SSIC 2020	SSIC 2020 Title
A	AGRICULTURE AND FISHING
01	AGRICULTURE AND RELATED SERVICE ACTIVITIES
011	GROWING OF CROPS, MARKET GARDENING AND HORTICULTURE
0111	<u>Growing of Food Crops (Non-Hydroponics)</u>
01111	Growing of leafy and fruit vegetables
01112	Growing of mushrooms
01113	Growing of root crops
01119	Growing of food crops (non-hydroponics) n.e.c.
0112	<u>Growing of Food Crops (Hydroponics)</u>
01120	Growing of leafy and fruit vegetables (hydroponics)
0113	<u>Growing of Fruits</u>
01130	Growing of fruits
0114	<u>Growing of Nursery Products</u>
01141	Growing of orchids
01142	Growing of ornamental plants
01149	Growing of nursery products n.e.c.
0119	<u>Growing of Other Crops</u>
01190	Growing of other crops
014	ANIMAL PRODUCTION
0141	<u>Livestock Production (except Poultry and Animal Specialties)</u>
01411	Pig farms
01412	Cattle farms (including dairy cattle)
01413	Goat farms (including goat's milk production)
0142	<u>Poultry Farms and Hatcheries</u>
01421	Poultry breeding/hatcheries
01422	Broiler farms (chickens reared for meat)
01423	Layer farms (chickens reared for eggs)
01424	Duck farms
0149	<u>Other Animal Production</u>
01491	Dog breeding

# Challenges III

Misalignment reconciliation and Context Fine Tuning

## Measure of Misalignment

- How misaligned the declared vs classification?
- Proxy metric based off rank and hierarchical levels
- Quality-Of-Life filter for investigation



## New business descriptions / SSIC codes

- Lack of annotated data from Entities
- Reference list from DOS
- SSIC code list refresh every 5 years
- Adhoc Fine-Tuning of distilBERT



SFA approves 16 insect species for food;  
companies gear up to offer new dishes  
and products



# Retrospective

Review for continuous improvement



## Type of Classification Approaches

- Text Classification via NLI vs FT-LM (vs RCS)
- RCS require larger annotated dataset
- Open-source vs Paid Subscription

## More Data sources = Better model?

- For model training: Quality > Quantity
- Importance of **Validated/Annotated** vs **Unvalidated/Unannotated** Data
- Web scraping ethics
- Extractive Summarization on targeted source section

## Legitimate source of Business Description

- Validated and Undisputable for classification
- Conglomerate, Multi-Industry/Sector Corporations subsidiaries

## Key questions

1. Any cost considerations, open vs close source models or solutions?
2. Which is the best source of **validated** data for SSIC definition reference? (model training & validation)
3. Which is the best source of **truth** data for Entity SSIC classification?  
Not SSIC declaration (model testing)
  - Specific page, section, paragraph (targeted)

# Next exploration?

Proposed distilBERT + Gemini Pro



DOS

Dictionary list of  
SSIC and Descriptions  
(Validated)



AR

Annual Report PDF  
(Unvalidated)



Web scrap

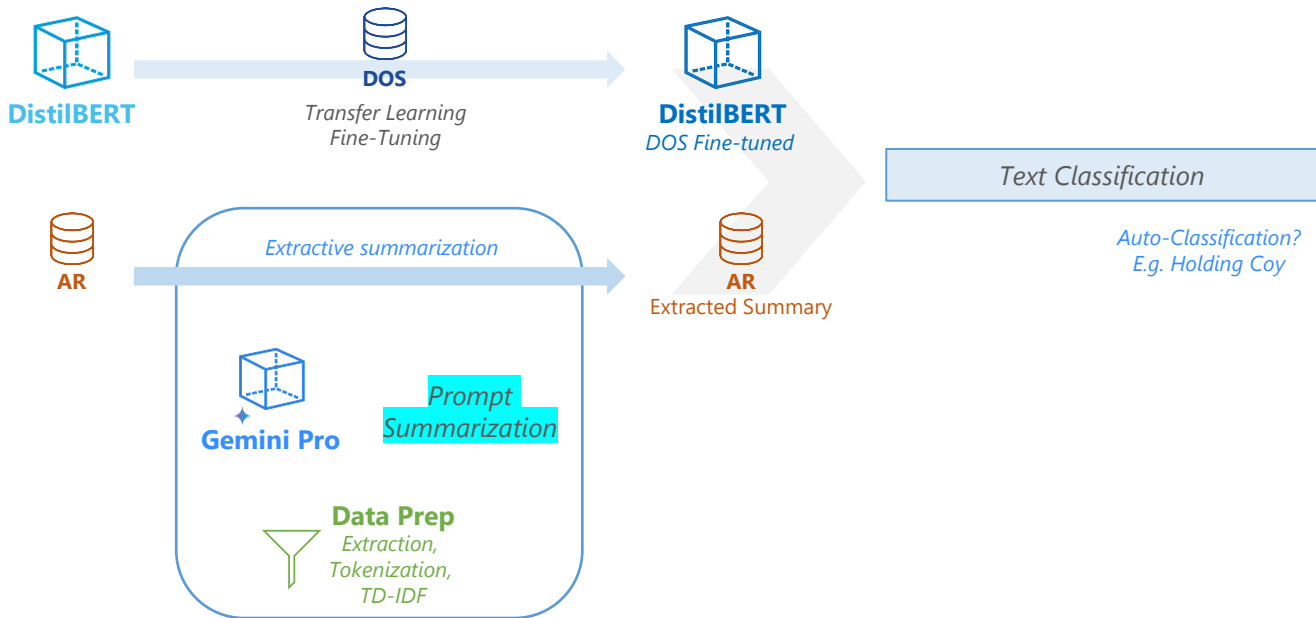
Web scrap LinkedIn &  
Corporate Website  
(Unvalidated)

**Not  
used**

## Proposed Exploratory Model

Fine-Tuning DistilBERT for Multiclass Text Classification

5 Different Models for each Level





# QnA

ssicsync

## Ang Mei Chi

Lew Kuan Teng **Roy**

Liu Wudi

# Michael Wong Wait Kit

Ong **Wee Yang**

