



Faculty of Computer Science and Information Technology

***A Study of FinBERT's Effectiveness in Financial Text Analysis: Insights into Sentiment, ESG Factors, and Predictive Outcomes from Malaysian Financial Institutions***

**FINAL YEAR PROJECT**

**HANNA ELLISA JOSEPH (81780)**

**BACHELOR OF COMPUTER SCIENCE WITH HONOURS (INFORMATION SYSTEMS)**

2024

**A STUDY OF FINBERT'S EFFECTIVENESS IN FINANCIAL TEXT  
ANALYSIS: INSIGHTS INTO SENTIMENT, ESG FACTORS, AND  
PREDICTIVE OUTCOMES FROM MALAYSIAN FINANCIAL  
INSTITUTIONS**

HANNA ELLISA JOSEPH

This project is submitted in partial fulfilment of the requirements for the degree of Bachelor  
of Computer Science with Honours (Information Systems)

**Faculty of Computer Science and Information Technology**  
**UNIVERSITI MALAYSIA SARAWAK**

**2024**

**KAJIAN KEBERKESANAN FINBERT DALAM ANALISIS TEKS KEWANGAAN:  
WAWASAN MENGENAI ANALISIS SENTIMEN, FAKTOR ESG, DAN HASIL  
PREDIKTIF DARIPADA INSTITUSI KEWANGAN MALAYSIA**

HANNA ELLISA JOSEPH

Projek ini merupakan salah satu keperluan Ijazah Sarjana Muda Sains Komputer dengan  
Kepujian (Sistem Maklumat)

**Fakulti Sains Komputer dan Teknologi Maklumat  
UNIVERSITI MALAYSIA SARAWAK**

**2024**

# Declaration

I hereby declare that this study, inclusive of all related content is the result of my own work.  
All sources of information used, including references and citations, have been duly acknowledged.

Signed,

A handwritten signature in black ink, appearing to read 'Hanna Ellisa Joseph', with a stylized flourish at the end.

Hanna Ellisa Joseph

81780

Faculty of Computer Science and Information Technology

University of Malaysia Sarawak

10 July 2025

# Acknowledgment

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Ts. Dr. Sarah Flora Samson Juan, for accepting me as one of her final year project students. Her invaluable expertise, support and guidance have been instrumental in the successful completion of this research. I would also like to thank my uncle, Mr. Phillip Wong, for his generous assistance and contribution to the progress of this study. I am also deeply grateful to my aunt, Dr. Mary Wong, for her constant encouragement and insightful advice throughout this journey. To my family — thank you for your endless support, patience and encouragement. To my friends, Macarena Douglas, Jeminah Paune, Eireanne Tirol, and the Attachment Issues group chat — I love all of you dearly. Thank you for listening to my rants and breakdowns and for reminding me that I am never alone. To the girls of Sublot 11/12, thank you for the warmth, laughter and moments of peace your company has brought me. Your companionship has meant more than words can say. To a special eight — thank you for keeping me company through countless late nights and for the comfort your voices have brought me. As you once said:

*“I’ll give you everything, just keep dreaming”*

And lastly, to my biggest cheerleader and critic — me — thank you for not giving up.

Thank you to everyone who contributed directly or indirectly to the successful completion of this work.

# **Abstract**

As the economy and financial landscape evolve, the ability to process and interpret large volumes of unstructured financial information becomes increasingly critical for both institutions and individual investors. This study aims to evaluate the effectiveness of the FinBERT, a specialised NLP model fine-tuned for financial text analysis, in analysing financial texts to extract sentiment, evaluate Environmental, Social, and Governance (ESG) factors and identify forward-looking statements (FLS) from corporate annual reports of Malaysian financial institutions. A comprehensive methodology, that includes a literature review, data collection, text preprocessing, the application of the FinBERT model to conduct financial text analysis, and fine-tuning the model, is employed. The findings of this study aim to assess the model's performance and its potential to enhance financial decision-making and bridge the gap between financial literacy and practical application.

**Keywords: Environmental, Social, and Governance (ESG), Forward-looking Statements (FLS), FinBERT, financial text analysis, Sentiment analysis**

## Abstrak

Apabila ekonomi dan landskap kewangan berkembang, keupayaan untuk memproses dan mentafsir sejumlah besar maklumat kewangan yang tidak berstruktur menjadi semakin kritikal untuk institusi dan pelabur individu. Penyelidikan ini bertujuan untuk menilai keberkesanan FinBERT, satu model NLP khusus yang ditala khas untuk analisis teks kewangan, menganalisis teks kewangan bagi mengekstrak sentimen, menilai faktor Alam Sekitar, Sosial, dan Tadbir Urus (ESG), serta mengenal pasti kenyataan berwawasan masa depan (FLS) daripada laporan tahunan korporat institusi kewangan di Malaysia. Satu metodologi yang komprehensif, yang merangkumi tinjauan literatur, pengumpulan data, prapemprosesan teks, penggunaan model FinBERT untuk menjalankan analisis teks kewangan, dan penalaan halus model, digunakan. Penemuan kajian ini bertujuan untuk menilai prestasi model serta potensinya untuk mempertingkatkan keupayaan membuat keputusan kewangan dan merapatkan jurang antara literasi kewangan dan aplikasi praktikal.

**Kata kunci:** analisis teks kewangan, analisis sentimen, FinBERT, kenyataan berpandangan ke hadapan (FLS), Sekitar, Sosial dan Tadbir Urus (ESG)

# Table of Contents

|                             |     |
|-----------------------------|-----|
| Acknowledgment .....        | i   |
| Abstract.....               | ii  |
| Abstrak.....                | iii |
| List of Figures .....       | vii |
| List of Tables .....        | x   |
| List of Abbreviations ..... | xii |

## Chapter 1: Introduction

|      |                             |    |
|------|-----------------------------|----|
| 1.1  | Introduction .....          | 1  |
| 1.2  | Problem Statement.....      | 3  |
| 1.3  | Aim and Objectives .....    | 4  |
| 1.4  | Research Methodology .....  | 4  |
| 1.5  | Project Scope .....         | 6  |
| 1.6  | Significance of Report..... | 6  |
| 1.7  | Project Schedule .....      | 7  |
| 1.8  | Expected Outcome.....       | 10 |
| 1.9  | Project Outline.....        | 10 |
| 1.10 | Chapter Summary .....       | 12 |

## Chapter 2: Literature Review

|         |                                                                            |    |
|---------|----------------------------------------------------------------------------|----|
| 2.1     | Introduction .....                                                         | 13 |
| 2.2     | AI in the Financial Sector .....                                           | 14 |
| 2.2.1   | Sentiment Analysis and its Impact on Financial Decision-Making.....        | 15 |
| 2.2.2   | Environmental, Social, and Governance (ESG) Evaluation in Finance.....     | 17 |
| 2.2.3   | Predictive Analytics and Forward-Looking Statements (FLS).....             | 18 |
| 2.3     | Natural Language Processing (NLP) and Large Language Models (LLMs).....    | 21 |
| 2.4     | Financial Capability: Bridging the Gap Between Literacy and Practice ..... | 25 |
| 2.5     | NLP Models for Finance .....                                               | 27 |
| 2.5.1   | FinBERT .....                                                              | 28 |
| 2.5.2   | BloombergGPT .....                                                         | 35 |
| 2.5.3   | FinGPT.....                                                                | 39 |
| 2.5.4   | Comparative Summary of FinBERT, BloombergGPT and FinGPT .....              | 41 |
| 2.5.5   | Techniques for Improving NLP Model Training.....                           | 42 |
| 2.5.5.1 | Class Weighting .....                                                      | 42 |
| 2.5.5.2 | Early Stopping Callbacks .....                                             | 43 |
| 2.6     | The Malaysian Financial Landscape: History and Context.....                | 44 |
| 2.6.1   | FinBERT for Financial Insights in Malaysian Bank Annual Reports.....       | 46 |
| 2.7     | Chapter Summary .....                                                      | 49 |

## **Chapter 3: Research Methodology**

|       |                                    |    |
|-------|------------------------------------|----|
| 3.1   | Introduction .....                 | 50 |
| 3.2   | Research Methodology .....         | 50 |
| 3.2.1 | Literature Review .....            | 52 |
| 3.2.2 | Data Collection .....              | 52 |
| 3.2.3 | Text Proprocessing .....           | 55 |
| 3.2.4 | Testing FinBERT Algorithm .....    | 58 |
| 3.2.5 | Fine-Tuning the FinBERT Model..... | 60 |
| 3.2.6 | Evaluation and Results .....       | 62 |
| 3.3   | Chapter Summary .....              | 63 |

## **Chapter 4: Testing and Fine-Tuning the FinBERT Model**

|         |                                            |    |
|---------|--------------------------------------------|----|
| 4.1     | Introduction .....                         | 64 |
| 4.2     | Dataset Composition .....                  | 64 |
| 4.2.1   | Structure of Dataset .....                 | 64 |
| 4.2.2   | Statistical Analysis of Dataset .....      | 68 |
| 4.2.2.1 | Word Frequency Analysis.....               | 68 |
| 4.2.3   | Test Set Composition.....                  | 72 |
| 4.3     | Experimental Setup .....                   | 75 |
| 4.3.1   | Testing Parameters.....                    | 75 |
| 4.3.2   | Tools and Libraries Used.....              | 75 |
| 4.3.3   | Testing Workflow .....                     | 76 |
| 4.4     | Fine-Tuning Workflow.....                  | 82 |
| 4.4.1   | Training Data Selection .....              | 82 |
| 4.4.2   | Parameters.....                            | 84 |
| 4.4.3   | Model Adaption and Training.....           | 85 |
| 4.5     | Model Deployment via Gradio Interface..... | 91 |
| 4.6     | Chapter Summary .....                      | 96 |

## **Chapter 5: Performance Evaluation and Results Analysis**

|       |                                                               |     |
|-------|---------------------------------------------------------------|-----|
| 5.1   | Introduction .....                                            | 97  |
| 5.2   | Model Training Overview .....                                 | 97  |
| 5.3   | Performance Evaluation (Pre Fine-Tuning).....                 | 101 |
| 5.3.1 | Sentiment Analysis Results .....                              | 101 |
| 5.3.2 | ESG Classification Results .....                              | 103 |
| 5.3.3 | Forward-Looking Statements (FLS) Classification Results ..... | 106 |
| 5.4   | Performance Evaluation (Post Fine-Tuning).....                | 108 |
| 5.4.1 | Sentiment Analysis Results .....                              | 108 |
| 5.4.2 | ESG Classification Results .....                              | 110 |

|                                                |                                                                 |     |
|------------------------------------------------|-----------------------------------------------------------------|-----|
| 5.4.3                                          | Forward-Looking Statements (FLS) Classification Results .....   | 113 |
| 5.5                                            | Analysis of FinBERT Model Performance.....                      | 115 |
| 5.5.1                                          | Comparative Analysis of Pre- and Post-Fine-Tuning Results ..... | 115 |
| 5.5.2                                          | Model Strengths and Weaknesses.....                             | 118 |
| 5.5.3                                          | Qualitative Prediction Examples .....                           | 120 |
| 5.5.4                                          | Error Analysis.....                                             | 126 |
| 5.5.5                                          | Limitations of the Evaluation Approach.....                     | 128 |
| 5.5.6                                          | Real-World Implications for Financial Text Analysis.....        | 129 |
| 5.6                                            | Discussion of Findings .....                                    | 131 |
| 5.7                                            | Chapter Summary .....                                           | 133 |
| <b>Chapter 6: Conclusion and Future Work</b>   |                                                                 |     |
| 6.1                                            | Introduction .....                                              | 134 |
| 6.2                                            | Achievements .....                                              | 134 |
| 6.3                                            | Constraints.....                                                | 137 |
| 6.4                                            | Future Work and Recommendations .....                           | 137 |
| 6.5                                            | Chapter Summary .....                                           | 138 |
| 6.6                                            | Project Conclusions .....                                       | 138 |
| References.....                                |                                                                 | 139 |
| Appendix A: Annotator Appointment Letter ..... |                                                                 | 145 |
| Appendix B: Annotator Form .....               |                                                                 | 146 |

## List of Figures

|                                                                                                       |    |
|-------------------------------------------------------------------------------------------------------|----|
| <b>Figure 1.</b> Phases of Research Methodology .....                                                 | 4  |
| <b>Figure 2.</b> Simplified Architecture of a Unidirectional RNN, adapted from Cui et al. (2018)..... | 28 |
| <b>Figure 3.</b> Overall pre-training and fine-tuning process for BERT. ....                          | 29 |
| <b>Figure 4.</b> Framework of FinGPT, as sourced from Yang et al. (2023).....                         | 39 |
| <b>Figure 5.</b> Workflow of Research Methodology .....                                               | 50 |
| <b>Figure 6.</b> Flowchart of the Data Collection Process .....                                       | 52 |
| <b>Figure 7.</b> Sample of Annual Report PDF .....                                                    | 54 |
| <b>Figure 8.</b> Flowchart of Text Preprocessing Process.....                                         | 56 |
| <b>Figure 9.</b> Sample of Segmented Sentences .....                                                  | 56 |
| <b>Figure 10.</b> Sample of Excel Corpus Structure for Sentiment Analysis.....                        | 57 |
| <b>Figure 11.</b> Sample of Excel Corpus Structure for ESG Classification .....                       | 57 |
| <b>Figure 12.</b> Sample of Excel Corpus Structure for Forward Looking Statement.....                 | 58 |
| <b>Figure 13.</b> Overview of Fine-tuning Process .....                                               | 61 |
| <b>Figure 14.</b> Percentage of Sentences by Institution .....                                        | 66 |
| <b>Figure 15.</b> Structure of Metadata .....                                                         | 66 |
| <b>Figure 16.</b> Code Snippet of Word Frequency Analysis and Word Cloud Generation.....              | 69 |
| <b>Figure 17.</b> Word Cloud of Most Frequently Used Terms in the Dataset.....                        | 71 |
| <b>Figure 18.</b> Structure of Test Set .....                                                         | 73 |
| <b>Figure 19.</b> Test Set Structure for Sentiment Analysis.....                                      | 74 |
| <b>Figure 20.</b> Test Set Structure for ESG Classification .....                                     | 74 |
| <b>Figure 21.</b> Test Set Structure for FLS Classification .....                                     | 74 |
| <b>Figure 22.</b> Workflow Overview .....                                                             | 76 |
| <b>Figure 23.</b> Code Snippet for Task Preparation (Sentiment Analysis).....                         | 77 |

|                                                                                                                |    |
|----------------------------------------------------------------------------------------------------------------|----|
| <b>Figure 24.</b> Code Snippet for Sentiment Analysis.....                                                     | 77 |
| <b>Figure 25.</b> Code Snippet for ESG Classification .....                                                    | 78 |
| <b>Figure 26.</b> Code Snippet for FLS Classification .....                                                    | 78 |
| <b>Figure 27.</b> Code Snippet for Pipeline Execution (Sentiment Analysis).....                                | 78 |
| <b>Figure 28.</b> Code Snippet for Batch Classification of Test Set Sentences .....                            | 79 |
| <b>Figure 29.</b> Code Snippet for Dataframe Conversion.....                                                   | 79 |
| <b>Figure 30.</b> Code Snippet for Performance Metrics Library .....                                           | 79 |
| <b>Figure 31.</b> Code Snippet for Importing and Preparing Prediction Results for Evaluation .....             | 80 |
| <b>Figure 32.</b> Structure of Excel File .....                                                                | 80 |
| <b>Figure 33.</b> Label Cleaning and Filtering .....                                                           | 80 |
| <b>Figure 34.</b> Code Snippet for Performance Metrics Evaluation and Confusion Matrix Plotting .....          | 81 |
| <b>Figure 35.</b> Breakdown of Sentence Contributions by Institution and Year .....                            | 83 |
| <b>Figure 36.</b> Code Snippet for Fine-Tuning Parameters (FinBERT FLS) .....                                  | 84 |
| <b>Figure 37.</b> Code Snippet for Libraries .....                                                             | 86 |
| <b>Figure 38.</b> Code Snippet for Model, Class Weights and Dataset Loading .....                              | 86 |
| <b>Figure 39.</b> Code Snippet for Custom BERT Subclass Implementation .....                                   | 87 |
| <b>Figure 40.</b> Code Snippet for BERT Tokenizer .....                                                        | 88 |
| <b>Figure 41.</b> Code Snippet for Trainer Initialisation with Early Stopping.....                             | 88 |
| <b>Figure 42.</b> Executing the Training Loop .....                                                            | 89 |
| <b>Figure 43.</b> Code Snippet for Model Evaluation (Post-Fine Tuning).....                                    | 89 |
| <b>Figure 44.</b> Code Snippet for Classification Report and Confusion Matrix Plotting (Post Fine-Tuning)..... | 90 |
| <b>Figure 45.</b> Code Snippet for Library Installation and Environment Setup .....                            | 91 |
| <b>Figure 46.</b> CPU Configuration for Model Inference.....                                                   | 92 |

|                                                                                                                  |     |
|------------------------------------------------------------------------------------------------------------------|-----|
| <b>Figure 47.</b> Loading Fine-Tuned FinBERT Models and Tokenizer from Hugging Face .....                        | 92  |
| <b>Figure 48.</b> Label Mapping and Classification Function Definition .....                                     | 92  |
| <b>Figure 49.</b> Code Snippet for Single Sentence Processing .....                                              | 93  |
| <b>Figure 50.</b> Code Snippet for Excel File Processing .....                                                   | 93  |
| <b>Figure 51.</b> Code Snippet for Gradio Interface Build.....                                                   | 94  |
| <b>Figure 52.</b> Training and Validation Loss with Accuracy Across Epochs for Sentiment<br>Analysis Model ..... | 98  |
| <b>Figure 53.</b> Training and Validation Loss with Accuracy Across Epochs for ESG<br>Classification Model ..... | 99  |
| <b>Figure 54.</b> Training and Validation Loss with Accuracy Across Epochs for FLS<br>Classification Model ..... | 100 |
| <b>Figure 55.</b> Confusion Matrix for Sentiment Analysis Classification.....                                    | 102 |
| <b>Figure 56.</b> Confusion Matrix for ESG Classification .....                                                  | 104 |
| <b>Figure 57.</b> Confusion Matrix for FLS Classification .....                                                  | 107 |
| <b>Figure 58.</b> Confusion Matrix for Sentiment Analysis Post Fine-Tuning.....                                  | 109 |
| <b>Figure 59.</b> Confusion Matrix for ESG Classification Post Fine-Tuning.....                                  | 112 |
| <b>Figure 60.</b> Confusion Matrix for FLS Classification Post Fine-Tuning .....                                 | 114 |
| <b>Figure 61.</b> Sentiment Analysis: Pre vs Post Fine-Tuning Performance.....                                   | 117 |
| <b>Figure 62.</b> ESG Classification: Pre vs Post Fine-Tuning Performance .....                                  | 117 |
| <b>Figure 63.</b> FLS Classification (Non-Specific FLS Class): Pre vs Post Fine-Tuning<br>Performance .....      | 118 |

## List of Tables

|                                                                                            |     |
|--------------------------------------------------------------------------------------------|-----|
| <b>Table 1.</b> Timeline for Final Year Project .....                                      | 7   |
| <b>Table 2.</b> Classification Results of Forward-Looking Statements in 10-K Reports.....  | 19  |
| <b>Table 3.</b> Variants of BERT and their key features .....                              | 30  |
| <b>Table 4.</b> List of BERT and FinBERT models. ....                                      | 32  |
| <b>Table 5.</b> Simplified breakdown of full training set used to train BloombergGPT.....  | 36  |
| <b>Table 6.</b> Comparison summary of FinBERT, BloombergGPT, and FinGPT. ....              | 41  |
| <b>Table 7.</b> Toolkit for Financial Text Analysis Using FinBERT .....                    | 47  |
| <b>Table 8.</b> Python Library Code and Their Purpose.....                                 | 54  |
| <b>Table 9.</b> Library Overview .....                                                     | 59  |
| <b>Table 10.</b> Code Snippet of Fine-tuned FinBERT Models.....                            | 60  |
| <b>Table 11.</b> Distribution of Sentences by Report .....                                 | 65  |
| <b>Table 12.</b> Distribution of Gold Standard Annotations .....                           | 67  |
| <b>Table 13.</b> Top 20 Most Frequent Terms.....                                           | 71  |
| <b>Table 14.</b> Tools and Libraries .....                                                 | 76  |
| <b>Table 15.</b> Dataset Composition for Model Fine-Tuning.....                            | 82  |
| <b>Table 16.</b> Training Output for Sentiment Analysis Model.....                         | 98  |
| <b>Table 17.</b> Training Output for ESG Classification Model .....                        | 99  |
| <b>Table 18.</b> Training Output for FLS Classification Model.....                         | 100 |
| <b>Table 19.</b> Classification Report for Sentiment Analysis .....                        | 101 |
| <b>Table 20.</b> Classification Report for ESG Classification.....                         | 103 |
| <b>Table 21.</b> Sample of Misclassifications by FinBERT model for ESG Classification..... | 105 |
| <b>Table 22.</b> Classification Report for FLS Classification. ....                        | 106 |
| <b>Table 23.</b> Classification Report for Sentiment Analysis Post Fine-Tuning. ....       | 108 |

|                                                                                                           |     |
|-----------------------------------------------------------------------------------------------------------|-----|
| <b>Table 24.</b> Classification Report for ESG Classification Post Fine-Tuning.....                       | 110 |
| <b>Table 25.</b> Classification Report for FLS Classification Post Fine-Tuning. ....                      | 113 |
| <b>Table 26.</b> Summary of Metric Improvements Across Tasks Post Fine-Tuning.....                        | 115 |
| <b>Table 27.</b> Summary of Model Strengths and Weaknesses by Classification Task (Post Fine-Tuning)..... | 119 |
| <b>Table 28.</b> Sentiment Analysis Predictions.....                                                      | 121 |
| <b>Table 29.</b> ESG Classification Predictions .....                                                     | 123 |
| <b>Table 30.</b> FLS Classification Predictions .....                                                     | 124 |
| <b>Table 31.</b> Summary of Research Questions .....                                                      | 134 |
| <b>Table 32.</b> Summary of Research Objectives .....                                                     | 135 |

## List of Abbreviations

| Abbreviation | Definition                                                        |
|--------------|-------------------------------------------------------------------|
| AI           | Artificial Intelligence                                           |
| BNM          | Bank Negara Malaysia                                              |
| BERT         | Bidirectional Encoder Representations From Transformers           |
| CIMB         | Commerce International Merchant Bankers Berhad                    |
| COVID-19     | Coronavirus Disease 2019                                          |
| ESG          | Environmental, Social, And Governance                             |
| FOMC         | Federal Open Market Committee                                     |
| FED          | Federal Reserve System                                            |
| FinBERT      | Financial Bidirectional Encoder Representations From Transformers |
| FLOPs        | Floating-Point Operations                                         |
| FLS          | Forward-Looking Statements                                        |
| LLMs         | Large Language Models                                             |
| LSTM         | Long Short-Term Memory                                            |
| MAYBANK      | Malayan Banking Berhad                                            |
| MFTC         | Malaysian Financial Text Corpus                                   |
| NLP          | Natural Language Processing                                       |
| SA           | Sentiment Analysis                                                |
| SVM          | Support Vector Machine                                            |

# Chapter 1: Introduction

## 1.1 Introduction

When Alan Turing first proposed ‘*Imitation Game*’, he laid the foundation for what would eventually form the basis for Artificial Intelligence (AI). The game illustrated a vision where machine responses were indistinguishable from that of humans, sparking early debates about the nature and potential of machine intelligence. Decades later, breakthroughs in advanced digital technologies and applications of artificial intelligence have fuelled remarkable technological progress (Bahoo et al., 2024). Today, AI has become a driving force across multiple industries and has transformed the way we interact with and analyse data. In the financial sector, where vast amounts of unstructured information – such as corporate reports, earnings calls and news – continuously shape market behaviour, AI-powered tools are essential in uncovering patterns and insights effectively. One such use of AI in the financial industry is financial text analysis, which transforms how institutions and individuals interpret financial information. According to Loughran & McDonald (2020), large-scale textual analysis has become a vital part of the methodological toolkit in finance, with Natural Language Processing (NLP) serving as its core.

Huang et al. (2022) discusses the introduction of NLP algorithms based on deep learning models. According to the study, these algorithms, more commonly referred to as large language models (LLMs), as a consequence of the extraordinarily large number of parameters, which can amount to several billion, “learn semantic and syntactic relations among words from a large volume of texts, which allows them to consider context (e.g., other words in the same text and word sequences) in summarising texts” (p.2). LLMs such as the FinBERT, a specialized variant of Google’s BERT (Bidirectional Encoder Representations from Transformers) model customised for financial text processing represent advancements in

Natural Language Processing (NLP) as it is pre-trained on large-scale financial datasets, giving it a comprehensive understanding of financial terminology, linguistic patterns, context-specific jargon, and further fine-tuned for specific tasks related to the financial industry. Examples of tasks include Sentiment Analysis (SA) which captures the emotional tone of market reports and news and provides knowledge regarding investor sentiment that could potentially influence market behaviour; Environmental, Social, and Governance (ESG) analysis, which evaluates companies' non-financial performance, helping stakeholders assess risks tied to sustainability and ethical practices; and Forward-Looking Statements (FLS) which focuses on forecasts, risks, and future strategies, offering critical insights into a company's strategic direction and market outlook.

However, beyond its role in institutional analysis, LLMs also contributes to enhancing public financial capability. According to the World Bank Group (2021), financial capability represents the ability to act on acquired financial knowledge (literacy), skills and behaviours to make informed decisions about managing personal resources and building financial resilience. It goes beyond understanding financial concepts, focusing on the practical use of that knowledge to achieve personal and financial well-being. As financial reports and news become more accessible, the general public (herein referred to as 'investors') engage more with complex financial terminology texts. A survey conducted by RinggitPlus this year found that only 38% of Malaysians believe they will only survive 3 months or less if they lose their job according to their current financial savings and lifestyle. This suggests an inability to manage resources effectively and build sufficient savings for emergencies. Although respondents may possess some level of financial literacy, they may struggle to apply that knowledge into practice, thus highlighting the gap between understanding financial concepts and being able to act on them to ensure financial resilience. LLMs help bridge this gap by summarizing key points, simplifying jargon, and offering insights in a more accessible manner (Head et al., 2023).

Therefore, this study examines the effectiveness of FinBERT in conducting financial text analysis, particularly in its ability to extract sentiment, evaluate ESG factors, and generate predictive outcomes from corporate annual reports of Malaysian financial institutions. As the financial landscape continues to evolve, tools like FinBERT play important roles in bridging gaps of knowledge between financial concepts and practical application, thereby contributing to greater financial capability and resilience among investors. By employing advanced NLP techniques, this study aims to highlight how FinBERT can equip stakeholders with the necessary knowledge and skills, ultimately fostering financially capable Malaysians.

## **1.2 Problem Statement**

Despite the large-scale implementation of AI in the last decade across multiple areas of the financial industry (Li et al., 2023), financial institutions and investors still encounter challenges in processing and interpreting large volumes of unstructured financial information. While models like FinBERT are able to complete tasks relevant to the field, questions remain regarding its effectiveness within specific contexts, such as the Malaysian financial landscape. Furthermore, a constant gap between knowledge, skill and practical application suggests the need for tools that not only enhance financial analysis but also empower users to make informed decisions. This study seeks to address these challenges by evaluating FinBERT's ability to analyse financial texts and explore how such tools can support financial institutions and investors in improving financial capability and resilience.

Below are the research questions this study will focus on:

1. How effective is FinBERT in analysing financial texts to extract sentiment, evaluating ESG factors, and generating predictive outcomes from corporate reports in the Malaysian financial sector?

2. How can FinBERT support improved financial decision-making for both institutions and individual investors?
3. How can FinBERT help bridge the gap between financial literacy and practical application for investors and the general public?

### 1.3 Aim and Objectives

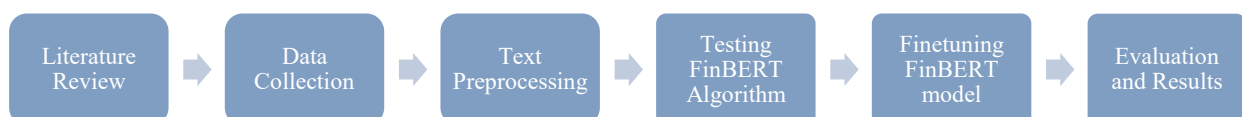
The aim of this study is to investigate the potential of advanced AI model, in particular the FinBERT model, in uncovering valuable insights from corporate financial texts of Malaysian financial institutions.

The primary objectives of this study are:

1. To review the fundamentals of LLMs and FinBERT, financial issues, and the role of AI in finance.
2. To design and develop an approach for utilising the fine-tuned FinBERT model to conduct sentiment analysis, ESG evaluation, and predictive analytics on financial texts.
3. To analyse and evaluate the outcomes of the FinBERT model in providing insights corporate financial reports in the context of sentiment analysis, ESG evaluation, and predictive analytics on financial texts.

### 1.4 Research Methodology

**Figure 1.**  
*Phases of Research Methodology*



The research methodology consists of five phases, beginning with a comprehensive review of relevant literature to gain insights into existing research related to the FinBERT model, its applications as well as the current financial landscape in Malaysia.

This is followed by the data collection phase where relevant corporate annual reports from various local financial institutions were identified. The reports spanned selected years (2019, 2021, 2023), offering a detailed view of financial trends and developments. 2019 represents both the pre-COVID-19 period and the early phase of the pandemic, which provides a look into the country's financial landscape at the onset of the pandemic. 2021 reflects the post-COVID-19 recovery phase, highlighting changes and adaptations in the financial sector following the pandemic's peak. Finally, 2023 captures the most recent developments, providing a current perspective on evolving financial trends and practices. Textual content from relevant sections of these reports was extracted using the PyPDF library, while non-textual elements like graphs and images were excluded. The extracted text was then organised into an Excel spreadsheet to be further processed.

Prior to the testing phase, the collected data is prepared to ensure compatibility with the FinBERT model. This involves segmenting the extracted text into individual sentences, organising them into a structured format, and manually annotating each sentence for analysis. The labelled sentences are compiled into a structured corpus and a final review of the corpus ensures the quality, consistency, and alignment of the data with the requirements of the three tasks.

With the pre-processed text prepared, the FinBERT model is employed to conduct several types of analysis. Sentiment analysis identifies the emotional tone of the reports, classifying text into positive, negative, or neutral sentiment categories. The model also deployed for ESG evaluation, detecting relevant environmental, social, and governance content

using specific keywords or phrases to assess company performance on these criteria. Additionally, FLS classification is conducted, identifying and analysing forecasts, risk assessments, and future strategies outlined in the reports.

Once testing is complete, the pretrained FinBERT model is fine-tuned using a subset of the corpus to adapt it to the nuances of Malaysian financial texts. The dataset is split into training, validation, and test subsets, and used to train the model on for sentiment analysis, ESG classification and FLS classification. The model's performance is evaluated and the fine-tuned model is tested for generalization on unseen data and saved for future use.

Lastly, the performance of the pretrained and fine-tuned FinBERT models is evaluated using a combination of accuracy and a confusion matrix to assess their effectiveness. Accuracy provides an overall measure of correct predictions, while the confusion matrix offers a detailed breakdown of prediction outcomes, including true positives, true negatives, false positives, and false negatives. Together, these metrics provide a comprehensive framework for understanding model performance and identifying areas for improvement.

## **1.5 Project Scope**

This project will focus on analysing corporate annual reports from Malaysian financial institutes, namely Bank Negara Malaysia (BNM), Malayan Banking Berhad (Maybank) and Commerce International Merchant Bankers Berhad (CIMB), covering the years 2019, 2021 and 2023. The study will employ FinBERT to conduct sentiment analysis, identify ESG factors, and interpret FLS, with an emphasis on how these analyses can enhance financial capability among Malaysian investors.

## **1.6 Significance of Report**

By evaluating the effectiveness of FinBERT in analysing financial texts within a local context, this study contributes to the growing field of NLP and its applications in finance. The

findings will offer valuable insights into the relationships among sentiment, ESG factors, and forward-looking statements, which are essential for informed investment decisions. This research aims to deepen the understanding of how LLMs operate in financial contexts. Utilising FinBERT will help develop practical skills in implementing NLP algorithms, processing unstructured data, and fine-tuning models for specific tasks. Additionally, the study highlights the broader significance of LLMs in the financial sector, particularly in extracting insights from complex documents like corporate reports. The results will clarify FinBERT's capability when applied to financial texts from Malaysian institutions, enhancing understanding of its effectiveness in sentiment analysis, ESG evaluation, and predictive analytics. Ultimately, these insights will illustrate how NLP tools can support financial decision-making and enhance financial capability for both institutions and individual investors in Malaysia.

## 1.7 Project Schedule

The project will be conducted over two semesters, namely Semester 1 2024/2025 and Semester 2 2024/2025. The duration of the project is from October 2024 to July 2025. Table 1 outlines the project schedule.

**Table 1.**  
*Timeline for Final Year Project*

| Details                                        | Start Date | End Date   | Duration |
|------------------------------------------------|------------|------------|----------|
| Development and Submission of Brief Proposal   | 9/10/2024  | 16/10/2024 | 8 days   |
| Feedback and Commentary from Reviewer/Examiner | 20/10/2024 | 28/10/2024 | 9 days   |
| Development and Submission of Full Proposal    | 4/11/2024  | 14/11/2024 | 10 days  |

|                                                                           |            |            |         |
|---------------------------------------------------------------------------|------------|------------|---------|
| Development and Submission of Chapter 1                                   | 4/11/2024  | 21/11/2024 | 17 days |
| Development and Submission of Chapter 2                                   | 22/11/2024 | 13/12/2024 | 21 days |
| Development and Submission of Chapter 3                                   | 28/12/2024 | 8/1/2025   | 12 days |
| Development and Submission of FYP 1 Final Report and Paper for assessment | 28/12/2024 | 15/1/2025  | 19 days |
| Development and Submission of Pre-Recorded Presentation Video and Slides  | 2/1/2025   | 19/1/2025  | 17 days |
| Evaluation Process through Pre-Recorded video by Supervisor and Examiner  | 19/1/2025  | 25/1/2025  | 7 days  |
| Amendment and Modification Period for FYP (if any)                        | 25/1/2025  | 6/2/2025   | 13 days |
| Submission of Amended Final Report (FYP 1)                                | 6/2/2025   | 17/2/2025  | 12 days |
| Dataset Preparation and Annotation                                        | 17/03/2025 | 17/4/2025  | 30 days |
| Initial Testing of Pretrained FinBERT Model                               | 1/4/2025   | 8/4/2025   | 7 days  |

|                                              |           |           |         |
|----------------------------------------------|-----------|-----------|---------|
| Evaluation of Pretrained FinBERT Model       | 8/4/2025  | 17/4/2024 | 10 days |
| Fine-Tuning FinBERT on Custom Dataset        | 18/4/2025 | 10/5/2025 | 23 days |
| Evaluation of Fine-Tuned FinBERT Model       | 11/5/2025 | 16/5/2025 | 6 days  |
| Development of Gradio Interface              | 18/4/2025 | 16/5/2025 | 30 days |
| Development and Submission of Chapter 4      | 8/4/2024  | 16/5/2025 | 44 days |
| Development of Chapter 5                     | 17/5/2025 | 31/5/2025 | 15 days |
| Development of Chapter 6                     | 25/5/2025 | 31/5/2025 | 7 days  |
| Final FYP Report Compilation                 | 3/6/2025  | 22/6/2025 | 20 days |
| Submission of Final Report, Paper and Poster | 22/6/2025 | 23/6/2025 | 2 days  |
| FYP Symposium                                | 30/6/2025 | 1/7/2025  | 2 days  |
| Submission of Amended Final Report           | 2/7/2025  | 28/7/2025 | 26 days |

## **1.8 Expected Outcome**

Below is a list of the expected outcomes of this study:

1. The FinBERT model classifies annual reports from Malaysian banks with a higher proportion of positive sentiment which is more likely to correspond with strong financial performance (e.g., increased profitability or positive stock price trends).
2. The FinBERT model analyses the generated output to understand how the data relates to or contains ESG classifications, highlighting trends in environmental sustainability, social responsibility, and governance practices within corporate reports.
3. The model identifies the number of times future investments and strategic growth plans are mentioned within the reports as predictors of future revenue growth or business expansion.

## **1.9 Project Outline**

### Chapter 1: Introduction

This chapter provides a brief overview of the study as well as a short introduction to the FinBERT model and the following chapters of the study.

### Chapter 2: Literature Review

The second chapter dives into related literature on the applications of AI in the financial sector, focusing on sentiment analysis, ESG evaluation, and predictive analytics. It also looks at the role of NLP and LLMs in finance, the definition of financial capability, the history of the Malaysian financial landscape and the specific tools utilised in this study.

### Chapter 3: Reserach Methodology

This chapter outlines the methodologies and processes used in the study, explaining their relevance, how they contribute to the research, and how they support the achievement of the project's objectives.

### Chapter 4: Testing and Fine-Tuning the FinBERT Model

The fourth chapter outlines the how the specified methodologies will be applied to the development of the implementation processes for analysing financial reports using FinBERT. Initially, the pretrained model is tested on a labelled test set to evaluate its baseline performance. Following this, the model is fine-tuned using a custom annotated training set. The fine-tuning process aims to increase the model's accuracy across all three classification tasks. Once the model is fine-tuned and redeployed, the result from both the pretrained and fine-tuned procedures are collected for further evaluation and analysis.

### Chapter 5: Evaluation and Results Analysis

The fifth chapter presents the evaluation of the FinBERT model's performance before and after fine-tuning across all three classification tasks. Using the labelled test set, the results are compared and improvements in the performance metrics are highlighted. An in-depth analysis of the classification outputs, key findings and patterns observed in the results is included. Furthermore, an error analysis is conducted to identify limitations and potential areas for improvement.

### Chapter 6: Conclusion and Future Work

The sixth chapter concludes the study by summarising the findings and affirming the research objectives. It highlights the key contributions of using FinBERT to conduct the three tasks. The chapter also discusses the limitations of the current approach and outlines potential

improvements. Finally, the project is formally closed, and recommendations for future research are provided.

## **1.10 Chapter Summary**

This chapter outlines the structure of the research project, beginning with an introduction to the significance of AI and the FinBERT model in financial text analysis. It presents the problem statement, detailing the challenges faced by financial institutions and investors, along with the research questions and objectives. The methodology describes the approach for data collection and analysis, while the scope and significance highlight the relevance of the study. Finally, the expected outcomes are discussed, such as the number of times future investments and strategic growth plans are mentioned within the reports being identified, with the literature review to follow in the next chapter.

## **Chapter 2: Literature Review**

### **2.1 Introduction**

This chapter presents a review of relevant literature related to this study, with a focus on the role of Artificial Intelligence (AI) in the financial sector. The main objective is to explore how AI, particularly Natural Language Processing (NLP) and Large Language Models (LLMs), is transforming key areas within the financial industry. Therefore, it is essential to examine previous literature regarding the technology, methodologies, and their applications. The review begins with examining essential tasks within the industry: sentiment analysis and its impact on decision making; the role of Environmental, Social, and Governance (ESG) evaluation in finance; and forward-looking statements (FLS) and their applications in the financial sector. In addition, this chapter also discusses the broader applications of NLP and LLMs in various fields, such as education and healthcare, as well as the potential benefits and drawbacks of deploying these technologies.

Following this, the focus shifts to financial capability and the gap between financial literacy and practical financial resilience. The definitions and components of financial capability – knowledge, skills and behaviours – are explored as well as the role of LLMs in bridging the gap between knowledge and real-world application. The review also examines the BERT model, FinBERT and other specialised NLP Models, namely BloombergGPT and FinGPT, along with their respective roles, advancements and applications within the financial industry.

This chapter also aims to discuss the current financial landscape of Malaysia and the unique challenges it faces, particularly regarding financial reporting and text analysis. While Malaysia's financial industry is well-regulated, with both conventional and Islamic finance sectors, the challenge lies within the complexity and accessibility of financial information.

Annual reports, which are key sources of this information, are often vast, detailed, and hundreds of pages long. This presents an opportunity for LLMs, such as the FinBERT model, to be integrated into the business process to efficiently summarise, analyse, and extract key data from these dense reports, making critical information more accessible and understandable to a wider audience.

## **2.2 AI in the Financial Sector**

As digital technology advances and AI theory and technology continue to mature, the roles of businesses, the characteristics of value creation, and even economic structures rapidly evolve as a result (Li et al., 2021). Amongst the many sectors impacted by these advancements is the financial industry. The digitisation of financial services has become a primary focus, with various institutions now offering digital platforms and mobile applications for convenient access to their range of services. This digital shift has opened avenues for AI-driven innovations, such as product and service personalisation, risk management and fraud detection. According to Kruse et al. (2019), the rise in AI-driven approaches is a result of the growing volume of “big data”, which demands innovative data analytics methods to extract value for businesses. Kruse et. al. also states that these innovations are crucial for financial institutions to enhance customer loyalty and satisfaction, as they allow for more tailored services and improved decision-making processes in an increasingly competitive landscape. As AI-driven innovations continue to shape the financial landscape, various tasks that were once complex and time-consuming have now become normalised in the industry. These include tasks such as sentiment analysis, ESG classification and predictive analytics. The following subsections delve deeper into these tasks and their applications on various areas of the financial industry.

### **2.2.1 Sentiment Analysis and its Impact on Financial Decision-Making**

Sentiment analysis, as defined by the Cambridge dictionary, is a method of computationally processing and analysing written content, particularly social media comments, to determine people's opinions or feelings about a topic. As discussed by Mishev et al. (2020), sentiment analysis is a key NLP task that benefits from advancements made in the field of language modelling, leading to improved results while becoming an increasingly vital medium for converting emotions and attitudes into valuable, actionable insights. The importance of sentiment analysis as a tool is further supported by Sohangir et al. (2018), who emphasises its growing role in addressing challenges concerning Big Data. According to the authors, “deep learning has this opportunity to provide a solution to address the learning and data analysis problem that exists in a massive amount of data (Big Data)”. In this context, due to significant hurdles found in domain adoption and streaming data that are still present in large-scale deep learning models, methods from sentiment analysis have grown in value as a means of obtaining significant insights that aid companies, organizations, and individuals in better utilising their Big Data.

In the financial industry, the way information is communicated to customers—whether through reports, advertisements, or customer service interactions—plays a crucial role in shaping their perceptions, decisions, and trust in financial institutions. The wording, clarity and tone can influence investment behaviours, risk assessments, and the overall relationship between financial organizations and their clients. Various news sources, such as company-specific, geopolitical, and macroeconomic issues, influence the financial market (Du et al., 2024). In a situation where market conditions are volatile and ever-changing, sentiment analysis emerges as a valuable asset for better understanding public perceptions, opinions, concerns and expectations. In a study conducted by Lak & Turetken (2017), the impact of sentiment analysis on consumer decision-making through online reviews was empirically

evaluated. It was found that sentiment scores, which quantify the emotional tone of reviews, can significantly enhance the efficiency of purchase decisions. The study was conducted in a controlled environment with 100 participants, who were tasked with assessing digital cameras by reviewing materials that either included sentiment scores, quantifying the emotional tone of the text, or did not feature these scores. It was revealed that the speed at which participants made purchasing decisions greatly increased when provided with the sentiment scores. This was also achieved with no notable decrease in their confidence regarding their choices, despite the boost in decision-making efficiency. In doing so, the study was successful in highlighting the role of sentiment analysis as a means for navigating overwhelming amounts of user-generated content more efficiently.

Further evidence from Loginova et al. (2021) through a study conducted on decision-making in cryptocurrency trading highlights the significance of sentiment analysis. The study, which employed aspect-based sentiment analysis (ABSA) demonstrated that investors were able to make more informed decisions based on nuanced insights with ABSA as compared to when traditional sentiment polarity and Latent Dirichlet Allocation (LDA) topics were used. Additionally, the results of the study indicated a significant correlation between the extracted topics and subsequent market movements. This suggests that certain topics and their associated sentiments can serve as indicators of future price changes, providing traders with actionable insights for their investment strategies.

The implications of these studies extend beyond individual consumer behaviour; they are also relevant for financial institutions and businesses. By integrating sentiment analysis into their communication strategies, these organizations can gain a deeper understanding of public sentiment regarding their products and services. This understanding can inform marketing strategies, enhance customer engagement, and ultimately influence investment behaviours and risk assessments. As financial institutions and businesses continue to adapt to the market

landscape, utilising sentiment analysis as a tool will be crucial in making data-driven decisions that align with consumer attitudes to ensure informed financial choices and lasting relationship.

### **2.2.2 Environmental, Social, and Governance (ESG) Evaluation in Finance**

The stakeholder theory proposed by Edward Freeman in 1984 explains the relationship between Corporate Social Responsibility (CSR) and stakeholder value. While shareholders are key stakeholders in an organisation, a firm should consider the interest of all its customers, suppliers, employees, investors, communities when performing their business activities as each group has interconnected relationships with the business. This theory highlights that businesses that align their operations with the needs and expectations of stakeholders can create long-term value and enhance their overall performance.

In recent years, the rising emphasis on ESG criteria has caused a notable transformation in the financial industry and investment market (Rane et al., 2024; Sipola et al., 2023). As the growing importance of sustainable and ethical issues continues to shape societal values, ESG considerations have gradually started to be included in investment strategies, corporate decision-making, and risk assessment processes (Chopra et al., 2024). Ahmad et al. (2023), which draws from Shakil et al. (2019), highlights that businesses that fail to incorporate ESG factors into their operational activities may face negative consequences as investors assess a company's social and environmental responses through its ESG performance. The authors also found that firms with low scores face high ESG disputes and this leads to negative effects on a company's reputation and financial performance.

Shakil et al. (2019) analysed the ESG performance data from 93 emerging market banks over the period from 2015 to 2018, found favourable correlations between environmental and social performance and the financial performance of banks in emerging economies. The link between these factors is suggested to be due to stakeholders' interest in

the company or bank ESG disclosure. Consequently, firms which engage in sustainable and socially responsible practices tend to perform better financially. In another study by Lee & Isa (2023), the impact of ESG practices on the financial performance of Shariah-compliant companies in Malaysia was investigated to determine if there was a difference in performance between ESG companies (companies that actively incorporate ESG practices into their operations and decision-making processes) and non-ESG companies. The study utilised ESG scores Refinitiv from the years 2010 to 2017 and focused on a sample of listed companies and measured the performance of the companies based on metrics such as return of assets (ROA) and Tobin's Q, a financial metric used to assess a company's market value relative to its asset replacement cost. The results of the study revealed that firms with higher ESG scores are perceived as more competitive and were likely to perform better as compared to firms with lower scores. This further supports the growing body of evidence emphasising the positive link between ESG practices and financial performance.

As the importance of ESG practices continues to grow, its importance in enhancing corporate strategies, strengthening stakeholder relationships, and driving financial success becomes increasingly clear. Companies that prioritise sustainable and socially responsible practices are better positioned to create long-term value and adapt to the evolving expectations of today's business landscape.

### **2.2.3 Predictive Analytics and Forward-Looking Statements (FLS)**

Predictive analytics is the use of statistical techniques, machine learning algorithms, and data mining to examine both recent and past data (Adesina et al., 2024) while Li et al. (2008) refers to FLS as “projections or expectations about future financial performance” which often incorporates assumptions and qualitative judgments. While FLS are typically speculative and subject to uncertainty, predictive analytics can support these statements by offering data-

driven forecasts, helping to improve their accuracy. The utilisation of both these methods in business and finance has changed the way companies assess future performance, risks, and opportunities.

Glodd et al. (2023) explored how the automated analysis of FLS could enhance stock predictions by improving the understanding of future company performance, thereby facilitating informed investment decisions. The study analysed 10-K reports from 80 randomly selected S&P 500 companies spanning the years 2006 and 2020, and utilised NLP techniques to identify FLSs within the reports. Three classifiers were trained for this purpose: Support Vector Machine (SVM) with a bag-of-words approach, Long Short-Term Memory (LSTM) neural networks, and BERT. The table below summarises the results of the study, highlighting the performance of each model based on metrics such as precision, recall, and F1-score:

**Table 2.**  
*Classification Results of Forward-Looking Statements in 10-K Reports.*

| Model | Precision |      | Recall |      | F1-score |      |
|-------|-----------|------|--------|------|----------|------|
|       | False     | True | False  | True | False    | True |
| SVM   | 0.90      | 0.67 | 0.78   | 0.84 | 0.84     | 0.74 |
| LSTM  | 0.85      | 0.61 | 0.75   | 0.75 | 0.80     | 0.67 |
| BERT  | 0.98      | 0.81 | 0.88   | 0.96 | 0.93     | 0.88 |

It was found that, among the three models, BERT achieved the highest precision (0.98), recall (0.88), and F1-score (0.93), making it the most effective tool for FLS classification. While the LSTM model identified more statements as FLS on average, BERT was the most conservative and accurate.

Beyond the methodologies of Glodd et al., other studies have also explored advanced techniques for market predictions, emphasising the integration of machine learning into the

process. One such study is Taskin & Akal (2024), which explored the impact of the COVID-19 crisis on communication strategies and consumer inflation expectations by analysing the Federal Open Market Committee (FOMC) minutes. To achieve this, the researchers utilised FinBERT, a domain specific adaptation of the BERT model, and its variant, FinBERT FLS, which was designed to detect FLS. The models were instrumental in identifying nuanced financial language and in assessing the forward-looking strength of the Federal Reserve System (FED)'s communications.

The results revealed FinBERT, in particular FinBERT FLS, was able to accurately capture semantic subtleties and tone of the FOMC minutes, revealing that while the FED maintained a consistent communication tone during the pandemic, the forward-looking strength of its statements remained unchanged. Through FLS classification, it was also revealed that the association between the FED's messaging and consumer inflation expectations weakened, suggesting a diminished influence of central bank communication on public expectations amidst the global crisis.

Through these studies, it can be observed that the integration of machine learning models significantly enhances the ability to analyse FLS. As the role of FLS in predicting market movements becomes more relevant, these models offer a more nuanced and accurate understanding of future financial performance. With the use of advanced algorithms, companies and investors can gain deeper insights into potential risks and opportunities, ultimately improving their decision-making and forecasting capabilities in an increasingly complex market environment.

## **2.3 Natural Language Processing (NLP) and Large Language Models (LLMs)**

NLP refers to the ability of computers to interpret, understand, and generate human language in a way that is meaningful (Khurana et al., 2023). Over the last few years, NLP, which is an intersection between linguistics and computer science, has seen significant advancements within the field (Webersinke, 2023). The main objectives of NLP, as defined in Khurana et al. (2023), revolves around several key activities that aim to facilitate the understanding and manipulation of human language by computers. These objectives include interpretation of Natural Language Data (NLD), which seeks to enable machines to understand human language by interpreting the context and meaning behind words and phrases; the analysis of language which is the breakdown of sentences into their components to understand grammatical structures, semantics, and syntax; and the manipulation data related to tasks like text generation, summarization, and translation.

NLP continues to play an important role in bringing human communication and machine understanding in the financial industry. However, as the field evolves, there arises a need to evaluate its potentials and limitations. Wang et al. (2024) discusses the advantages and disadvantages of NLP, listing a few key aspects in each category. Among the numerous applications of NLP, one notable advantage highlighted by the author was timely risk identification. Using NLP to quickly analyse extensive financial reports, news articles, and social media posts, firms can provide timely risk warnings and decision support. NLP techniques are also well-suited for the financial sector, as they can efficiently process and analyse vast amounts of data, which addresses the industry's significant data demands. Additionally, another notable advantage of NLP is its ability to achieve high accuracy in financial risk detection by capturing data patterns and correlations through deep learning models such as LSTM networks.

However, while these advantages offer significant benefits to the industry, NLP still has its share of disadvantages. Wang et al. (2024) discusses several disadvantages of NLP implementation in the financial industry including computational complexity. The authors state that in time-sensitive environments, the extended training times and increased computational complexity of models like SVM can present challenges. There are also issues concerning data imbalances where the performance of traditional models such as the Random Forest model may be affected especially when dealing with financial datasets where data is often highly imbalanced. The effectiveness of NLP model performance is also heavily reliant on high quality input data and this is where poorly structured data may lead to potential drawbacks.

Building on the foundations of NLP, LLMs, like the BERT and GPT models, represent a significant leaps in the ability of machines to process and generate human-like text. These models are statistical tools trained on extensive text corpora to predict the probability distribution of word sequences. Zhao et al. (2024) states that the capabilities of LLMs to analyse extensive text data positions them as a valuable tool in the financial sector and discusses the various applications of LLMs. One example is the ability to generate information regarding market movements, conduct risk assessments and support investment choices by analysing financial reports, market updates and shareholder communications. Zhao et al. additionally notes another application of LLMs, which is the capability to handle natural language queries, delivering prompt financial guidance and assistance. Several studies support this, such as Haque et al. (2022), Liu et al. (2023) and Wu et al. (2023). However, despite the growing potential of LLMs as tools to support decision making processes, LLMs also come with certain advantages and disadvantages that need to be considered as these considerations play critical roles in determining the effectiveness and application of LLMs across various domains.

Nie et al. (2024) explores the integration of LLMs within the financial industry and considers several advantages. According to the author, LLMs possess an advanced understanding of financial terminology and industry-specific jargon and thus, are able to accurately interpret complex financial documents and news article, which enhances the accuracy of tasks such as sentiment analysis. Chu et al. (2023) supports this perspective by highlighting that LLMs, through Natural Language Understanding capabilities, “are adept at comprehending domain-specific terminology and generating human-like text”. Another advantage of LLMs is the ability to adapt to specific tasks through transfer learning. LLMs are initially pre-trained on vast corpora consisting of various topics and languages before implementation and can be fine-tuned with minimal domain-specific data to perform specialised tasks. This reduces their dependency on large, domain-specific dataset which allows it to be adapted efficiently for other tasks.

While LLMs offer significant benefits, their limitations should not be overlooked. The integration of LLMs in the financial industry should be ethically evaluated as highlighted by Nie et al. (2024). The author stresses that there may arise a bias in decision-making and a lack of clear accountability mechanisms for developers, which can lead to unintended negative outcomes if not properly addressed. A study by Cao et al. (2024) supports this by highlighting algorithmic accountability and transparency in financial decision-making. Ensuring transparency in the role of LLMs is essential for maintaining accountability and establishing trust between financial institutions and clients. Furthermore, biases inherent in the training data may perpetuate societal inequalities and lead to discriminatory outcomes, such as unequal access to financial services for specific demographics. Nie et. al also points out the potential for legal concerns regarding the usage of LLMs in finance, noting that liability and compliance with regulations may be called into question as the use of robot-generated content does not always meet the necessary standards or legal requirements. Despite this, LLMs have

demonstrated valuable applications across various fields, namely education and healthcare, and have shown their potential applicability beyond the financial sector.

Education revolves around the transfer of knowledge, real-time feedback and emotional interaction, with LLMs playing a key role in enhancing the immediacy of feedback in this process (Xu et al., 2024). The usage of LLMs in education has the ability to present tailored, adaptive learning experiences for students. Xu (2023) also highlights the capability of LLMs are to generate advanced mathematical reasoning, which aids the education sector in demonstrating their proficiency in multi-modal engagement, intelligent adaptive teaching, and self-supervision. LLMs also show great potential in the healthcare sector, where the need for efficient, personalised, and data-driven solutions is necessary. A study by Omar et al. (2024) explored how NLP and LLMs could be used for infectious disease diagnosis and prediction. The methodology included systematically reviewing major databases including PubMed, Embase, Web of Science, and Scopus, up to December 2023, in addition to employing the Quality Assessment of Diagnostic Accuracy Studies-2 tool for evaluating the quality and robustness of the included studies. It was found that LLM application, such as GPT-4, were able to detect urinary tract infections while BERTweet was efficient in the surveillance of Lyme Disease through social media analysis. Both application demonstrated effective disease monitoring and public health tracking capabilities. However, these models' effectiveness differed from study to study. While certain NLP tools were able to accurately detect invasive mold diseases and pneumonia in medical records, others performed less well, especially when it came to treating bloodstream infections. As NLP and LLMs technologies continue to advance, further refinement and exploration is needed to enhance their effectiveness in various domains, ensuring the reliability and accuracy of results.

## **2.4 Financial Capability: Bridging the Gap Between Literacy and Practice**

The World Bank Group defines financial capability as the “internal capacity to act in one's best financial interest, given socioeconomic environmental conditions” and the ability to act on acquired financial knowledge (literacy), skills and behaviours make informed decisions about understanding, choosing, and using financial services that meet personal needs, managing personal resources, and building financial resilience. Studies by Çera et al. (2021), Garg & Singh (2018), Potrich et al. (2015) and Santini et al. (2019) have all noted that financial capability is made up of three important components: financial attitude, financial knowledge, and financial behaviour. Kempson et al. (2013), as cited by Çera et al., claims that in contrast to the narrow notion of financial literacy, financial capability – consisting of both ability and opportunity – has been introduced to expand the concept. Individuals who are financially capable should possess both ability and opportunity to enhance their financial well-being through sound financial decisions and actions. However, if they possess the necessary skills and knowledge to be financially capable but fail to apply them in practical decision-making, they can be considered financially incapable (Piotrowska, 2019).

A study by Bialowolski et al. (2022) examined the role of financial literacy for financial resilience among middle-aged and older adults. By utilising a time-to-event analysis over a ten-year observation window, the authors attempted to investigate and understand how financial resilience can be preserved through the help of financial literacy. The results found that older adults, particularly those aged 65 and above, are more likely to score low on financial literacy assessments, which correlates with a higher risk of financial vulnerability. Statistically, 35.1% of individuals aged 75 and older scored low on financial literacy, compared to 14.9% of those under 65. It was also noted that financially literate adults aged 50 years and older had lower hazards of falling into a financially vulnerable state over a 10-year follow-up period, indicating that financial literacy contributes to maintaining financial stability over time. These findings

highlight the importance of understanding of financial concepts, regardless of age and socioeconomic status, to ensure informed and practical decision-making, financial stability, and resilience against potential economic challenges.

Mokhtar et al. (2020) studies four financial capability domains – planning ahead, managing money, choosing products and staying informed – and how these domains are affected by the influence of socio-demographic criteria such as age and ethnicity. The participants of this study comprised of 2000 respondents representing four major ethnic groups (Malay, Chinese, Indian, and Others) and four distinct demographic categories: rural residents from FELDA areas, private sector employees, government employees, and students in higher education institutions. The result revealed significant differences across age groups and their ability to be financially capable. Respondents aged 50 and above, which accounted for 18.4% of all respondents, had a higher mean score as compared to the other four age groups (below 20, 20-29, 30-39 and 40-49) at 20.42 which suggests that older individuals tend to manage their money more effectively than younger individuals, likely due to greater experience and knowledge in financial management practices. It was also noted that 55.3% of respondents saved less than 10% of their monthly income while around 11% of respondents reported that they did not save at all. Older respondents were also more likely to seek and engage with financial information, which helps to enhance their financial literacy and capability, as compared to the younger respondents. However, the study revealed that overall engagement with financial indicators and professional advice was low across all age groups, with fewer than half of the respondents consistently tracking financial indicators (46%) or consulting professional advisors (40%). Respondents also struggled to apply their knowledge effectively, when interpreting, reading, and comparing financial products, reflecting a lack of confidence in practical application of financial knowledge. The findings of this study clearly illustrate significant gaps in financial literacy and consequently, the practical application of financial

knowledge especially among younger people, highlighting the need for targeted initiatives to bridge these gaps.

One possible solution to address these gaps is through LLMs, which function as tools that can be used to simplify complex financial jargon and make financial information more accessible to diverse audiences. LLMs can break down complex, domain-specific terminology by simplifying intricate terms and concepts into more straightforward language (Zhao et al., 2024). This capability enhances the accessibility of financial information, especially for individuals with limited financial knowledge. Zhao et al. also highlights the capabilities of LLMs in offer explanations that account for the nuances of particular scenarios by accounting for the context in which financial terms are used. Additionally, the authors stated that by utilising NLP technology, LLMs are able to offer immediate guidance and support on user queries. This enables individuals to ask questions about financial topics without needing to understand the technical terminology, thus contributing to increasing user engagement and comprehension.

In conclusion, the integration of LLMs in finance not only simplifies complex information but also democratizes financial knowledge, thereby promoting inclusivity and allowing stakeholders to make informed financial decisions.

## **2.5 NLP Models for Finance**

The financial industry presents unique challenges for the implementation of NLP in its business processes due to its reliance on domain-specific tasks, technical jargon, and a rapidly evolving data and business environment. To address these challenges, specialised NLP models have emerged, designed specifically for financial tasks. These models include FinBERT, a domain-specific adaptation of BERT for financial text; BloombergGPT, a proprietary model leveraging Bloomberg's extensive dataset; and FinGPT, an open and flexible platform for

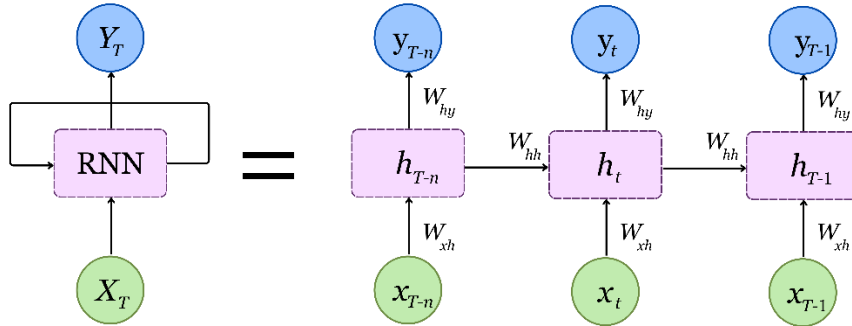
financial analysis. To understand how these models help address hurdles, their architectures, applications, and strengths must be studied.

### 2.5.1 FinBERT

Devlin et al. introduced the Bidirectional Encoder Representations from Transformers (BERT) model in 2019. It was developed to transform NLP by allowing machines to interpret the meaning of words in a sentence more accurately. Compared to earlier models that processed text in a unidirectional manner, BERT utilised a bidirectional approach by considering both the left and right contexts of each word simultaneously.

Before BERT, existing NLP models, such as Recurrent Neural Networks (RNNs), processed text left-to-right or right-to-left, as illustrated in the figure below:

**Figure 2.**  
*Simplified Architecture of a Unidirectional RNN, adapted from Cui et al. (2018)*

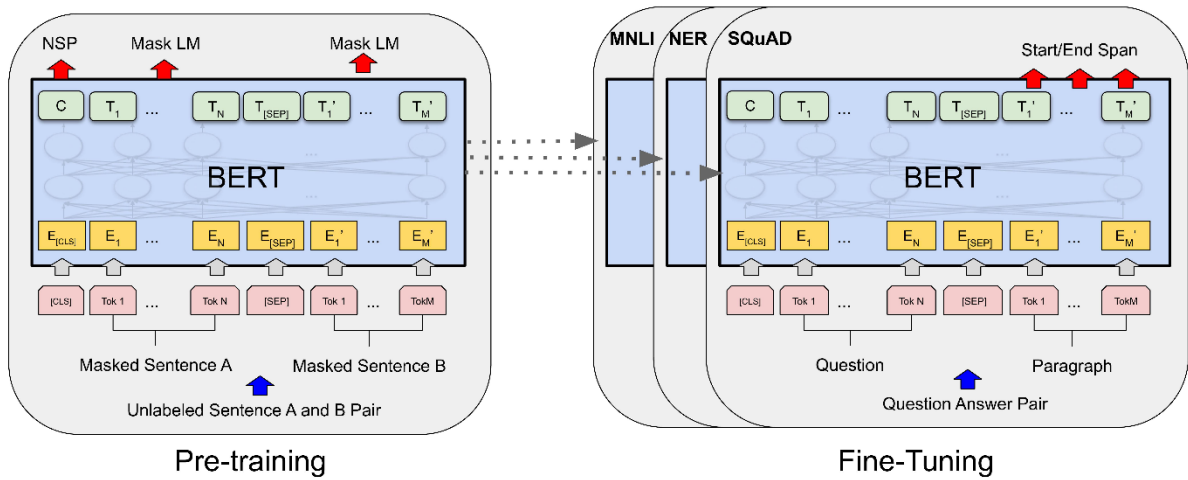


The model processes a sequence of inputs one at a time. Each hidden state ( $h$ ) is computed based on the current input and the previous hidden state, thus only capturing information provided by past inputs as the sequences progresses. The arrows indicate the flow of information, moving strictly from left to right. Outputs ( $y$ ) are generated at each time step ( $W_{hy}$ ) that map the hidden states to the output layer. As a result of the unidirectional processing, their ability to fully understand the contextual relationship between words is limited, particularly when analysing sentence-level or token-level tasks.

However, BERT overcame this limitation by pre-training on large corpora using two techniques: Masked Language Modeling (MLM), where a portion of the input tokens is masked and the model is trained to predict these tokens based on their context, and Next Sentence Prediction (NSP), where BERT is trained on pairs of sentences to predict whether the next sentence logically follows the first. This allowed the BERT model to address several challenges present in NLP, namely the limitations surrounding contextual understanding and task-specific processes which required models to undergo separate training and design for each NLP task. BERT's pre-training and fine-tuning process changed this approach as the model was pre-trained on vast, general-purpose datasets to learn universal language representations. These representations could then be fine-tuned with minimal task-specific data, thereby facilitating adaptation to diverse NLP applications.

**Figure 3.**

*Overall pre-training and fine-tuning process for BERT, as extracted Devlin et al. (2019).*



Since its introduction, several variants of BERT have been developed. These include:

**Table 3.**  
*Variants of BERT and their key features*

| Variant                 | Key Feature                                                     |
|-------------------------|-----------------------------------------------------------------|
| DistilBERT <sup>1</sup> | Smaller and faster variant of BERT.                             |
| RoBERTa <sup>2</sup>    | Removes NSP and trains on larger datasets for longer durations. |
| ALBERTA <sup>3</sup>    | Shares parameters across layers and decomposes embeddings.      |
| SciBERT <sup>4</sup>    | Pre-trained on scientific texts                                 |
| BioBERT <sup>5</sup>    | Pre-trained on biomedical texts.                                |

These variants demonstrate the adaptability of BERT’s architecture and its potential for domain-specific applications. While BERT excels at general NLP tasks, its application to specialised domains requires further adaptation. Specialised models have come about, such as SciBERT and BioBERT for the science and biomedical fields respectively. In the financial field, where precise language and context is important, FinBERT was introduced to address the linguistic challenges of financial text.

FinBERT (Financial Bidirectional Encoder Representations from Transformers), first introduced in 2019 by Araci, was built by further pre-training and adapting the BERT model for the financial domain using three main dataset: TRC2-financial, the Financial PhraseBank

<sup>1</sup> Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. <http://arxiv.org/abs/1910.01108>

<sup>2</sup> Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. <http://arxiv.org/abs/1907.11692>

<sup>3</sup> Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. <http://arxiv.org/abs/1909.11942>

<sup>4</sup> Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. <http://arxiv.org/abs/1903.10676>

<sup>5</sup> Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>

from Malo et al. (2013) and, FiQA, a dataset developed for WWW '18 conference financial opinion mining and question answering challenge. The TRC2-financial dataset was the result of filtering the original TRC2 dataset by Reuters into a corpus consisting of 46,143 documents of more than 29M words and 400k sentences. The resulting model was FinBERT.

To test the effectiveness of the model, Araci conducted a series of experiments that evaluated its performance on specific financial sentiment analysis tasks. The two primary datasets used were the FiQA Sentiment Dataset<sup>6</sup>, which was used to assess the model's performance in sentiment scoring and the Financial PhraseBank, which provided labelled financial sentences for sentiment classification tasks. The metrics of evaluation of the model's effectiveness included: Accuracy, which measured the proportion of correctly classified instances out of the total instances, providing a straightforward evaluation of classification performance; Macro F1 Average, which calculated the F1 score for each class and then averaged them and, as noted by Araci, is particularly useful in cases of label imbalance, as seen in the Financial PhraseBank dataset; Cross Entropy Loss quantified the difference between the predicted and actual probability distributions, aiding in performance evaluation during training and testing; Mean Squared Error (MSE) was used to evaluate the model's abilities for regression tasks with a focus on sentiment scoring; and R<sup>2</sup> Score assessed how well the model's predictions aligned with actual data points. The latter two metrics were used to evaluate the regression model as they were also standard measures commonly used to test the FiQA dataset (Araci, 2019). Additionally, to benchmark the model's capabilities, FinBERT was compared against other pre-trained models, namely ELMo (Embeddings from Language Model) and ULMFit (Universal Language Model Fine-Tuning), and the model was further validated using 10-fold cross-validation, a method that divides the dataset into ten subsets, iteratively training on nine subsets and testing on the remaining one to minimise dependency on a specific train-

---

<sup>6</sup> <https://sites.google.com/view/fiqa/home>

test split. Lastly, Araci utilised a confusion matrix to visualize the model’s classification performance, providing insights into the types of errors made (e.g., false positives and false negatives) and helping to identify areas for improvement.

The results of the study found that the FinBERT model showed that even when using a limited labeled dataset (as few as 500 training examples), pre-trained models can perform well on sentiment analysis tasks. This disproves the idea that deep learning models are unduly "data-hungry". In comparison, ULMFit, which was also further pre-trained on a financial corpus was not able to surpass FinBERT’s performance despite achieving favourable results. FinBERT also set new benchmarks on the FiQA sentiment scoring dataset and the Financial PhraseBank dataset as it improved classification accuracy by 15% over previous leading models. Overall, Araci (2019) highlighted FinBERT's effectiveness in financial sentiment analysis, showcasing its ability to utilise pre-trained language models for specialised tasks in the financial domain.

In a subsequent study of the FinBERT model, published in 2020 by Yang et al., new variants of the model were introduced. The variants, along with the BERT Base models used in this study are listed in the table below:

**Table 4.**  
*List of BERT and FinBERT models.*

| Model Name |                             |
|------------|-----------------------------|
| BERT       | BERT-Base (Cased)           |
|            | BERT-Base (Uncased)         |
| FinBERT    | FinBERT-BaseVocab (Cased)   |
|            | FinBERT-BaseVocab (Uncased) |
|            | FinBERT-FinVocab (Cased)    |

The term "cased" refers to models that distinguish between uppercase and lowercase letters and use the base vocabulary with cased tokenization, whereas "uncased" models do not make this distinction and use the base vocabulary with uncased tokenization.

The initial version, FinBERT Base-Vocab, underwent additional pretraining from a BERT Base checkpoint using three financial corpora: the Corporate Reports 10-K & 10-Q from the Securities Exchange Commission<sup>7</sup>, Earnings Call Transcripts from the Seeking Alpha website<sup>8</sup>, and Analyst Reports from the Investext database. The second version, FinBERT FinVocab, was trained from scratch on the same three corpora, but rather than using the vocabulary from the original BERT, Yang et al. used new vocabulary tailored specifically to the financial domain, none of which were inherent to the original BERT dataset. Each dataset was randomly split into 90% for training and 10% for testing, and this process was repeated 10 times to ensure robustness in the results. Three experiments were conducted comparing the original BERT-Base model (Devlin et al., 2019) against the FinBERT variants and the average scores was recorded. The metric for evaluating the effectiveness of the FinBERT model was accuracy. The study found that FinBERT consistently outperformed generic BERT models on three financial sentiment classification tasks, with the FinBERT-FinVocab model reporting the highest average score on each of the datasets used (Financial Phrase Bank: 0.872, FiQA Dataset: 0.844, AnalystTone Dataset: 0.887), thus proving the effectiveness of the FinBERT model in handling domain specific tasks.

Peng et al. (2021) also studied the effectiveness of FinBERT comparing its performances with the original BERT model. The study utilised the FinBERT models

---

<sup>7</sup> <https://www.sec.gov/edgar.shtml>.

<sup>8</sup> <https://seekingalpha.com/>

introduced by Y. Yang et al. (2020) to test the model's performance on a range of tasks, including sentiment analysis, document causality detection (Mariko et al., 2020), numeral understanding (Chen et al., 2019b) and numeral attachment (Chen et al., 2020). Several datasets were used as each task had different requirements. For sentiment analysis, three main datasets were used: the Financial Phrase Book (Malo et al., 2013), FinTextSen (Daudert et al., 2018), StockSen (Xing et al., 2020). For the document causality detection task, the authors utilised the dataset of the FinCausal shared task 2020 (Mariko et al., 2020), which comprises of texts from a 2019 corpus of financial news. Meanwhile, for the numeral understanding and attachment tasks, the FinNum-1 subtask  $\frac{1}{2}$  (Chen et al., 2019b) and FinNum-2 (Chen et al., 2019a) were respectively used. The study used two evaluation metrics: Macro F1-score (calculates the F1-score for each class independently and records the average, giving equal importance to each class regardless of size) and Micro F1-score (calculated by summing the true positives, false negatives, and false positives across all classes, making it more sensitive to class imbalances).

The results recorded by this study show that the FinBERT-BaseVocab model performed the best in almost all datasets. It outperformed the BERT Base model, with an average increase of 0.85 in the Macro F1-score across different benchmarks. It was also noted that the model excelled particularly in numeral-related tasks. The FinBERT FinVocab model, which was previously the best performing model in Yang et al. (2020), achieved the highest score in the numeral attachment and causality detection (Micro F1) tasks but otherwise recorded minimal increases in other tasks. This suggests that the choice of pretraining strategy can significantly impact model effectiveness depending on the specific task.

In terms of results and evaluation metrics, the three studies highlighted share a common approach of assessing FinBERT's performance on financial sentiment tasks, but they each expand the scope or introduce new evaluation strategies. All three studies employed Macro F1-score and accuracy as primary evaluation metrics for assessing FinBERT's effectiveness in

sentiment analysis and other financial tasks. While Araci (2019) demonstrated FinBERT's superiority over other models like ULMFit, Yang et al. (2020) and Peng et al. (2021) extended the evaluation to new financial tasks and found that FinBERT outperformed the original BERT, with domain-specific vocabulary boosting performance in tasks like numeral understanding and causality detection. The studies collectively indicate that the choice of pretraining strategy, particularly domain-specific vocabularies (FinVocab vs. BaseVocab), plays a crucial role in improving FinBERT's performance across diverse financial tasks. In conclusion, these findings highlight the significant impact of pretraining strategies, particularly the choice of vocabulary, on the model's performance across different financial tasks.

### **2.5.2 BloombergGPT**

BloombergGPT is a proprietary LLM developed by Bloomberg, introduced in a study by Wu et al. (2023), and designed specifically to support an extensive list of NLP tasks within the financial industry. The model is a decoder-only causal language model based on BLOOM (Scao et al., 2022), and was trained using a dataset taken from Bloomberg's own repository of financial text. Unlike general-purpose models, BloombergGPT takes advantage of its access to domain-specific data and vocabulary to ensure a high standard of performance on financial texts, capitalising on its ability to understand the nuances of financial communication.

The foundation of BloombergGPT's capabilities lies in FinPile – a curated dataset that includes both structured and unstructured financial data from Bloomberg's collection of financial documents, company filings, market data, and news, spanning the years 2007 to 2022. 51.27% of its training data comes from this dataset, which accounts for about 363 billion tokens. The remaining 48.73% is made up of publicly available datasets.

**Table 5.**  
*Simplified breakdown of full training set used to train BloombergGPT.*

| <b>Dataset</b> | <b>Average number of<br/>characters per document</b> | <b>Average number of<br/>characters per token</b> | <b>Percentage of overall<br/>tokens (%)</b> |
|----------------|------------------------------------------------------|---------------------------------------------------|---------------------------------------------|
| <b>FinPile</b> | 1,017                                                | 4.92                                              | 51.27                                       |
| Web            | 933                                                  | 4.96                                              | 42.01                                       |
| News           | 1,665                                                | 4.44                                              | 5.31                                        |
| Filings        | 2,340                                                | 5.39                                              | 2.04                                        |
| Press          | 3,443                                                | 5.06                                              | 1.21                                        |
| Bloomberg      | 758                                                  | 4.60                                              | 0.70                                        |
| <b>Public</b>  | 3,314                                                | 4.87                                              | 48.73                                       |
| C4             | 2,206                                                | 5.56                                              | 19.48                                       |
| Pile-CC        | 4,401                                                | 5.42                                              | 6.02                                        |
| GitHub         | 5,364                                                | 3.38                                              | 3.20                                        |
| Books3         | 552,398                                              | 4.97                                              | 3.02                                        |
| PubMed Central | 31,282                                               | 4.51                                              | 2.96                                        |
| ArXiv          | 47,819                                               | 3.56                                              | 2.35                                        |
| OpenWebText2   | 3,850                                                | 5.07                                              | 1.80                                        |
| FreeLaw        | 15,381                                               | 4.99                                              | 1.52                                        |
| StackExchange  | 2,201                                                | 4.17                                              | 1.15                                        |
| DM Mathematics | 8,193                                                | 1.92                                              | 0.60                                        |
| Wikipedia (en) | 2,988                                                | 4.65                                              | 0.53                                        |
| USPTO          | 4,339                                                | 6.18                                              | 0.51                                        |
| Backgrounds    |                                                      |                                                   |                                             |
| PubMed         | 1,333                                                | 5.77                                              | 0.50                                        |
| Abstracts      |                                                      |                                                   |                                             |

|                     |                     |                    |                      |
|---------------------|---------------------|--------------------|----------------------|
| OpenSubtitles       | 31,055              | 4.90               | 0.34                 |
| Gutenberg (PG-19)   | 399,351             | 4.89               | 0.32                 |
| Ubuntu IRC          | 539,222             | 3.16               | 0.25                 |
| EuroParl            | 65,053              | 2.93               | 0.21                 |
| YouTubeSubtitles    | 19,831              | 2.54               | 0.19                 |
| BookCorpus2         | 370,384             | 5.36               | 0.17                 |
| HackerNews          | 5,009               | 4.87               | 0.12                 |
| PhilPapers          | 74,827              | 4.21               | 0.08                 |
| NIH ExPorter        | 2,165               | 6.65               | 0.04                 |
| Enron Emails        | 1,882               | 3.90               | 0.02                 |
| Wikipedia           | 3,271               | 3.06               | 3.35                 |
| <b><i>TOTAL</i></b> | <b><i>1,531</i></b> | <b><i>4.89</i></b> | <b><i>100.00</i></b> |

The model also utilises a unique Unigram tokenizer as compared to conventional merge-based sub-word tokenizers which contributes significantly to its performance against other LLMs. Through its training framework, the model is able to excel in financial NLP tasks, such as sentiment analysis, named entity recognition, and finance-specific question answering, making it an effective tool for industry-use.

BloombergGPT has several industrial applications, as highlighted by Wu et al. Its ability to interpret and analyse financial jargon accurately makes it a valuable tool for professionals in investment analysis, risk management, and regulatory compliance as documents in these fields are usually nuanced. Investment bankers benefit from the capabilities of the model to process and summarise large volumes of financial data as well as analyse market trends, generate pitch books and prepare financial models, which contributes to the

improvement of decision making and client presentations. Asset managers can also use BloombergGPT to analyse portfolios, assess risks and report on the financial performance of their firms. With its advanced performance in sentiment analysis and financial question answering, the model serves as a valuable resource in assisting managers to make well-informed decisions using real-time market data and news. Another application of BloombergGPT noted by Wu et al. in the industry is in retail banking. The model can enhance customer service by powering chatbots and virtual assistants, thus providing customers with the ability to answer inquiries about products and services without the need for direct human intervention.

However, Li et al. (2023) states that while the BloombergGPT model performs well on industry-specific tasks, it still falls short in more complex tasks that require deeper semantic understanding and structured prediction compared to some fine-tuned models. The study, which explored and compared the performance of LLMs like ChatGPT and GPT-4 against specific fine-tuned models (FinBERT and BloombergGPT) also indicated that further research and improvements to the model are still needed before it can meet industry requirements. When compared to the performance of other models that were studied alongside, BloombergGPT's performance was competitive but not the best. The model achieved a weighted F1-score of 82.20 in the headline classification task, which was significantly lower than the F1-score of the FinBERT model (95.36). Additionally, in the sentiment analysis tasks, BloombergGPT was consistently outperformed by the FinBERT model, specifically when deep domain knowledge was required.

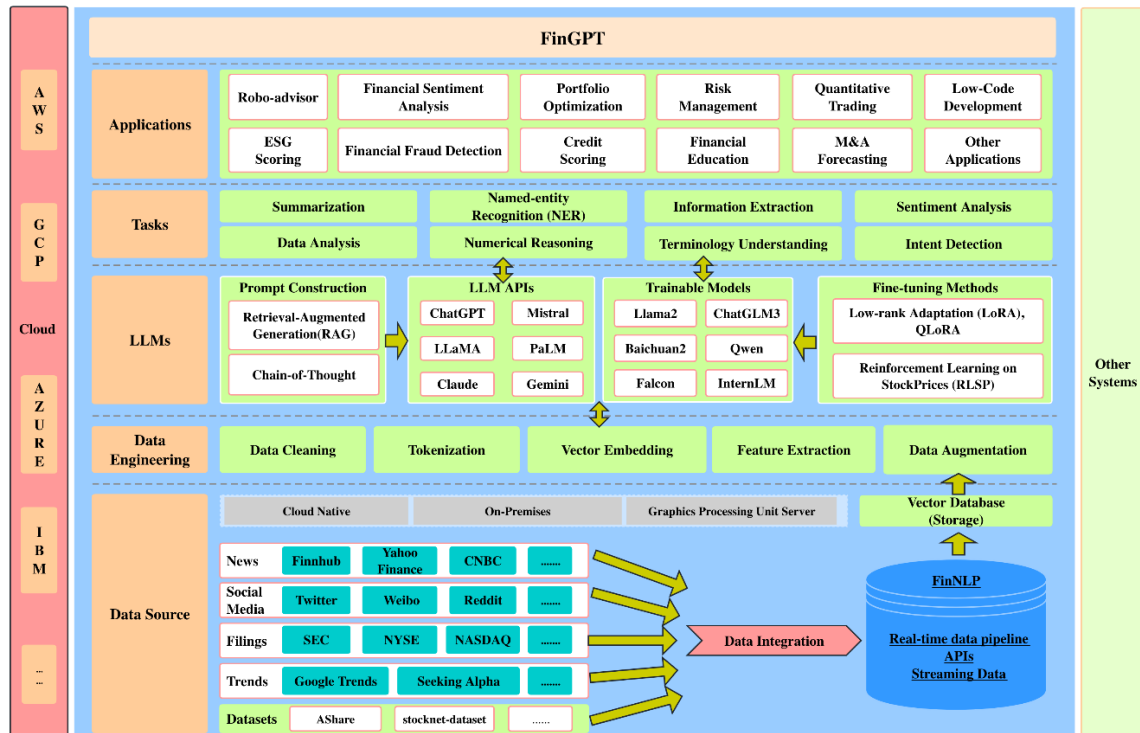
Overall, BloombergGPT has the potential to function as a specialised tool for financial application. While there is room for further development, its ability to handle financial analysis and deliver valuable insights positions it as a promising asset in the industry.

### 2.5.3 FinGPT

FinGPT is an open-source LLM designed as a response to the challenges associated with accessing high-quality financial data. The model was introduced by Yang et al. (2023) and is the product of efforts of researchers from Columbia University and New York University (Shanghai). The framework aims to democratize access to financial data and provide researchers and practitioners with the tools necessary to develop their financial language models.

Other than its open-source nature that promotes transparency and allows for user customisation, the developers of FinGPT took a data-centric approach by designing a framework that focused on obtaining, cleaning and preparing high-quality data.

**Figure 4.**  
*Framework of FinGPT, as sourced from Yang et al. (2023)*



FinGPT's framework is structured into four distinct layers: Data Source, Data Engineering, LLMs, and Application. The Data Source layer is responsible for acquiring

financial data from various sources such as financial news website, social media, financial statements and news articles. The second layer (Data Engineering) focuses on real-time processing of the acquired data. This layer is in charge of addressing hurdles within the processing stage such as high temporal sensitivity and low signal-to-noise ratio which are common challenges in financial data processing. NLP techniques are employed in this layer to filter noise and emphasise only the most relevant information. The third layer is the core of the framework and consists of a collection of fine-tuning LLMs. It employs lightweight adaptation techniques to ensure the model is consistently relevant and updated for the fast-paced environment of the industry. Lastly, the Application layer is designed to showcase the practical applications of the FinGPT model. This includes demonstrations and tutorials for using the model on financial tasks and serves as a guide for potential users.

Yang et al. (2023) highlights several industry applications of FinGPT. Among them are its ability for robo-advising, which is done through the development of automated financial advisory services. This provides users with personalised investment recommendations based on data and market conditions and helps improve decision making for individual and institutional investors. In a study by Liang et al. (2024), FinGPT was used to improve the accuracy of predictions regarding short-term stock price movements. The results of the study showed that there was an increase in prediction accuracy from the baseline accuracy of 55% to 63%. The study also reported that FinGPT also showed better performance at high clustering ratios (greater than 50%), which highlighted the importance of effective new clustering to improve prediction outcomes.

While FinGPT presents advantages in terms of accessibility, data handling and application versatility, it also faces several challenges. One such challenge is its dependency on high-quality labelled data. As noted by Yang et al. (2023), despite the emphasis on data quality, the model's effectiveness is still largely reliant on the availability of such data, which

can be time-consuming and costly to acquire. It is also observed that due to the inherent complexity of financial data, accurate modelling and prediction can become complicated.

#### 2.5.4 Comparative Summary of FinBERT, BloombergGPT and FinGPT

**Table 6.**

*Comparison summary of FinBERT, BloombergGPT, and FinGPT.*

| Model        | Strengths                                                                                | Limitations                                                                | Suitability for Study                                                  |
|--------------|------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|------------------------------------------------------------------------|
| FinBERT      | Specialises in NLP tasks (sentiment analysis, named entity recognition), high precision. | Limited scope compared to more versatile models.                           | Most suitable for tasks requiring high accuracy and domain expertise.  |
| BloombergGPT | Versatile, capable of a wide range of financial tasks using NLP.                         | Lacks fine-tuning for specialised tasks, restricted by proprietary nature. | Less suitable due to lack of fine-tuning for specific financial tasks. |
| FinGPT       | Open-source framework, flexible                                                          | Struggles with reliance on high-quality labelled data.                     | Less suitable due to data dependency and lack of domain precision.     |

FinBERT, BloombergGPT, and FinGPT each have their strengths and contribute differently to the financial language field. BloombergGPT, while versatile and capable of performing a wide range of financial tasks using NLP, lacks the fine-tuning required for specialised tasks and is restricted by its proprietary nature. FinGPT also struggles in certain aspects, in particular its reliance on high-quality labelled data despite its open-source framework.

In contrast, FinBERT stands out for its specialisation and precision when performing financial NLP tasks such as sentiment analysis and named entity recognition. Unlike the other two models, FinBERT is able to consistently produce reliable results in domain-specific tasks.

Given the focus of this study on tasks requiring high accuracy and domain expertise, FinBERT is the most suitable model to be utilised. The choice to use the model can be attributed to its ability to handle nuanced financial language to deliver consistent results which align with the objectives of this study, thus making it the ideal choice.

### **2.5.5 Techniques for Improving NLP Model Training**

In recent years, the fine-tuning of pretrained models has become an increasingly popular method to customise language models for domain-specific tasks. This approach, where large pretrained models like BERT, are adapted for specialised applications like financial classifications tasks, reduces the need to train models from scratch while improving task-specific performance.

However, fine-tuning also introduces several challenges that can compromise model performance if not properly addressed. Among the most common are class imbalances, where certain labels are significantly underrepresented in the training data, and overfitting, whereby the model learns patterns that do not generalise beyond the training set. To address these issues, class weighting and early stopping are usually implemented as precautionary steps.

#### **2.5.5.1 Class Weighting**

The concept of class weights was first formalised and implemented as part of the Scikit-learn framework by Pedregosa et al. (2011). Class weighting is used to handle imbalanced datasets, especially in situations where the majority class may be more favoured due to an unequal distribution of class labels, which may then cause biased model predictions.

To minimise these effects, Scikit-learn provides the `compute_class_weight` function, which automatically assigns a heavier weight to underrepresented classes and a lower weight to overrepresented classes, thereby creating a balance.

When `class_weight = 'balanced'`, the weights are calculated based on the following formula:

$$\frac{n_{samples}}{n_{classes} \times np_{bcount}(y)}$$

where:

- $n_{samples}$  is the total number of training samples
- $n_{classes}$  is the number of unique classes in the dataset
- $np_{bcount}(y)$  returns the number of occurrences (frequency) of each class label in the array ( $y$ )

Through this method, the model learn from all classes, regardless of how frequently they occur in the dataset, by ensuring that the loss function penalises any errors on the minority class more heavily. The loss function configuration, `CrossEntropyLoss(weight = class_weights)` from the PyTorch-based implementation module, then passes the calculated class weights through the model's training process.

#### 2.5.5.2 Early Stopping Callbacks

One common challenge when training deep learning models is overfitting, which is when the model learns the training data too well, including any noise or outliers that may be present, leading to poor performance and generalisation on new, unseen data. Overfitting is more likely to occur when working with limited or noisy domain-specific datasets and happens most during the fine-tuning process. In order to mitigate this is by introducing early stopping in the fine-tuning methodology, which halts the fine-tuning process if the performance on the validation set stops improving.

The `EarlyStoppingCallback` class is the part of the Hugging Face transformers library and works as a method to implement early stopping. This class serves as the plug-and-play `TrainerCallback` that monitors selected evaluation metrics, such as the validation set performance or accuracy, automatically during the training process and stops when improvements stagnate.

The mechanism for early stopping is implemented through two key parameters: `early_stopping_patience` (int), where int is the number of consecutive evaluation steps that can occur with no improvement in the metrics. If there is no progress within the set int number of steps, the training process is stopped; and `early_stopping_threshold` (float, optional), which indicates the marginal change, however small, for a metric to be considered an significant improvement. But anything below this threshold will not be counted as progress.

This mechanism is used in combination with the `TrainingArguments` parameter `load_best_model_at_end = True` which allows the model to return to the best performing checkpoint after stopping training. The trainer will need the same step interval for saving and evaluating the model (`save_steps = eval_steps`) to run early stopping behaviour correctly.

Early stopping is useful in financial NLP tasks as datasets may be imbalanced or sparse, which can often lead to degradation of performance and training beyond optimal capacity. Therefore, this method improves generalisation and cuts down on unnecessary computations by monitoring the performance of the validation set and halting training (if necessary) at the right time.

## **2.6 The Malaysian Financial Landscape: History and Context**

Since achieving independence in 1957, Malaysia has experienced a remarkable economic transformation, shifting from an economy largely reliant on agriculture and commodities to one driven by manufacturing and services. This shift has been supported by a

growing financial sector, with Bank Negara Malaysia (BNM) playing a central role as the nation's central bank and primary regulator. BNM has played a key role in facilitating the country's economic growth, driving investment, and supporting the transition to a more diversified and modern economy. Building on this foundation, the financial industry of Malaysia has diversified, with "27 conventional commercial banks and 16 full-fledged Islamic banks from both the local and foreign-owned entities" as of 2020 (Lui et al., 2020). Over the years, the financial landscape in Malaysia has transformed, allowing the sector to adapt to both regional and global challenges. This evolution has supported business expansion, fostered the development of new financial products, and increased international financial integration.

In recent years, Malaysia has also seen an increased focus on the digitalisation of banking services, with BNM introducing the Digital Banking Licensing Framework in 2019 and the Financial Sector Blueprint 2022-2026, designed to promote the adoption of technology and increase financial inclusion. The Digital Banking License in particular opened doors for the establishment of digital-only banks such as GXBank, AEON Bank, and Boost Bank. The heightened emphasis on digital banking solutions and technologies has led to the integration of AI, blockchain, and big data analytics in business processes to improve operational efficiency and customer experiences.

The evolving nature of Malaysia's financial landscape presents many opportunities for advanced technologies like NLP to be integrated in everyday business processes. In this context, FinBERT emerges as one such resource. The practical applications of FinBERT within the financial industry is noted by Huang et al. (2022), which states that the model shows considerable promise for professionals in finance, investment, and regulatory sectors because of its ability to effectively perform several tasks, namely sentiment analysis of various financial documents such as analyst reports, earnings conference calls, discussions regarding ESG issues and predicting market movements for future investments. As banks and financial institutions

generate vast amounts of unstructured textual data, such as annual reports, FinBERT can ease and streamline reporting and analysis. The FinBERT model's usability is not limited to its usage within institutions but also as a resource for individual investors to make informed decisions. It delivers important insights into investor sentiment, market trends, and how news and events could impact financial markets, helping to make informed decisions (Shen & Zhang, 2024).

Through past research, it can be concluded that FinBERT has shown favourable performance, consistently outperforming other models in financial tasks. Araci (2019) showed that FinBERT set new benchmarks on datasets like the Financial PhraseBank and FiQA, while its edge over conventional BERT models was demonstrated by Yang et al. (2020), with FinBERT-FinVocab achieving the highest accuracy across multiple financial sentiment classification tasks. Peng et al. (2021) further validated FinBERT's effectiveness, particularly in numeral understanding and causality detection. These findings reveal that pre-trained models like FinBERT perform well even with limited datasets, and pretraining strategies play an important role in its performance.

As Malaysia continues to embrace digital transformation in its financial sector, FinBERT's capabilities in tackling financial tasks present a valuable opportunity for application in the field, thereby aligning with the nation's vision for a digitally empowered economy.

### **2.6.1 FinBERT for Financial Insights in Malaysian Bank Annual Reports**

This study aims to measure the effectiveness of FinBERT's abilities to conduct sentiment analysis, ESG classification and forward-looking statement classification on annual reports from banks in Malaysia. The reports were chosen from three institutions: Bank Negara Malaysia (BNM), as the central bank, which provides a regulatory and macroeconomic

perspective into financial regulation and economic policies, and Malayan Banking Berhad (Maybank) and Commerce International Merchant Bankers Berhad (CIMB), which represents two of Malaysia's largest local banks, both of which offer a look into the commercial banking perspective. Additionally, the reports chosen consists of those from selected years: 2019, to represent both the pre-COVID-19 period and the early phase of the pandemic; 2021, which covers the post-COVID-19 recovery phase; and 2023, which provides the most recent developments and current perspective on trends and practices in the industry.

To carry out this study, an NLP toolkit will be utilised to preprocess, prepare and test FinBERT's effectiveness using the textual data extracted from the selected annual reports. Below is a table detailing the toolkit that will be utilised throughout this study.

**Table 7.**  
*Toolkit for Financial Text Analysis Using FinBERT*

| <b>Process</b>              | <b>Library</b>                                 | <b>Purpose</b>                                                        |
|-----------------------------|------------------------------------------------|-----------------------------------------------------------------------|
| Numerical and Data Handling | numpy, pandas                                  | Organises and preprocess text data for analysis.                      |
| Text Extraction             | PyPDF2                                         | Extracts text from PDF files.                                         |
| Data Preprocessing          | nlTK.tokenize,<br>nlTK.corpus.stopwords,string | Tokenizes, removes stopwords and cleans text.                         |
| Label Encoding              | sklearn.preprocessing.LabelEncoder             | Converts categorical labels into numerical format for model training. |
| Data Splitting              | sklearn.model_selection                        | Used to split dataset for training.                                   |

|                               |                                                                                                       |                                                                                     |
|-------------------------------|-------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| Model Loading and Fine-Tuning | transformers,torch,<br>BertTokenizer,<br>BertForSequenceClassification,<br>TrainingArguments,pipeline | Provides tools for pre-trained FinBERT model, fine-tuning and text classification.  |
| Performance Evaluation        | sklearn.metrics                                                                                       | Used to evaluate FinBERT model's performance.                                       |
| Visualisation                 | matplotlib, seaborn                                                                                   | Generates visual representations (i.e., confusion matrices, classification reports) |
| Interface Deployment          | gradio                                                                                                | Develops an interactive web-based interface to test and deploy the trained models.  |

The expectation of results from the FinBERT model is to extract meaningful insights from the annual reports of Malaysian financial institutions. Being a transformer-based model specifically pre-trained on financial texts, FinBERT is expected to accurately process and classify the nuanced language used in these reports. The following is the expectation of results by task:

#### 1. Sentiment Analysis

The primary expectation is that FinBERT will classify the text of the annual reports into distinct sentiment categories—positive, negative, and neutral. It is expected for FinBERT to provide a reliable understanding of the tone of the reports, revealing whether the overall sentiment of the institution is optimistic, cautious, or neutral.

## 2. ESG Classification

The FinBERT is expected to accurately identify and classify text related to Environmental, Social, and Governance (ESG) factors. It is expected that the model will be able to classify ESG-related disclosures in the reports, such as environmental sustainability efforts, corporate social responsibility initiatives, and governance practices.

## 3. FLS Classification

The third expectation is that FinBERT will effectively identify forward-looking statements in the annual reports. These statement often originate and contain predictions about future financial performance, market conditions, and strategic directions. This study expect FinBERT to be able to detect such forward-looking language, including projections, forecasts, and strategic goals outlined by financial institutions.

## 2.7 Chapter Summary

The focus of this chapter was to discuss several important topics, namely, the current advancements of AI in the financial sector, its applications and the state of the Malaysian financial landscape. Specifically, this chapter highlighted specialised NLP models such as FinBERT, BloombergGPT, and FinGPT, which are tailored to address the unique challenges of financial language and data analysis. The definition and concept of financial capability, its components, and how it relates to financial literacy and decision-making was also discussed in this chapter, along with the role of financial capability in promoting financial resilience. Additionally, the chapter also emphasised how Malaysia's evolving financial landscape presents opportunities for integrating advanced technologies like NLP to improve efficiency, transparency, and financial inclusion. Lastly, the components (the dataset, toolkit and expected outcome) of this study were introduced. The research methodology of this study will be presented in the subsequent chapter

# Chapter 3: Research Methodology

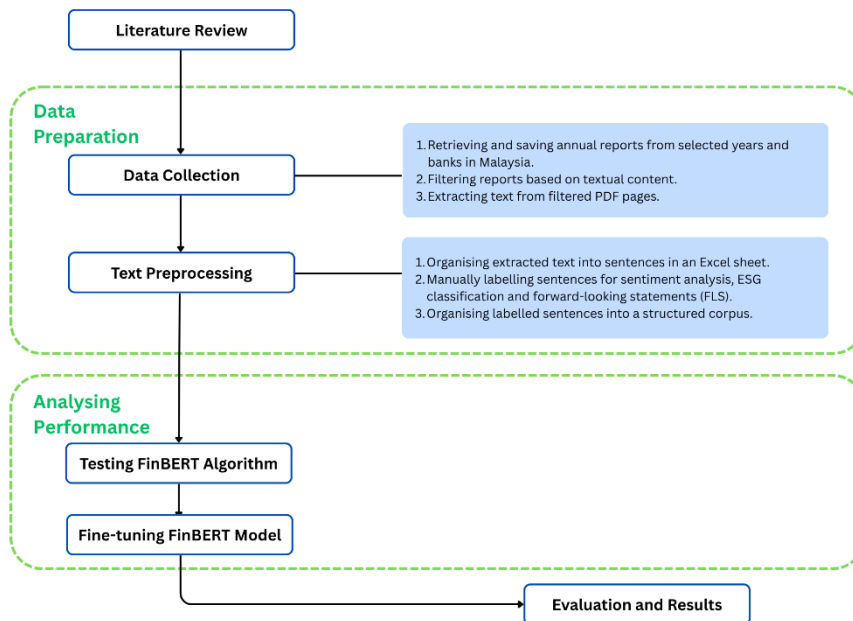
## 3.1 Introduction

The following chapter introduces the research methodology phases of this study. To ensure a systematic approach to this study, the project is divided into distinct phases, making it more organized and easier to comprehend.

## 3.2 Research Methodology

Fundamentally, there are six essential phases involved in this study, namely the literature review, data collection, text preprocessing, testing, fine-tuning the model as well as the evaluation and results phase.

**Figure 5.**  
*Workflow of Research Methodology*



The figure above details the research methodology of this study, which is divided into four main stages: Literature Review, Data Preparation, Analysing Performance, and Evaluation and Results. These stages are designed to ensure a systematic approach to data collection,

preprocessing, model testing and fine-tuning, and performance evaluation, ultimately achieving the research objectives with reliability and precision.

The Literature Review, conducted in Chapter 2, provided a comprehensive examination of academic and industry sources to establish the theoretical foundation of the study. This stage addresses Objective 1 by reviewing the fundamentals of LLMs, FinBERT, financial issues and the role of AI in finance, with a focus on their applications in the Malaysian financial sector.

Following this is the Data Preparation stage, which involves collecting annual reports from selected Malaysian banks, filtering the reports for relevant textual pages and content, and extracting the text for analysis. This stage primarily supports Objective 2, as it sets the foundation for using the fine-tuned FinBERT model for sentiment analysis, ESG evaluation, and forward-looking statement classification on financial texts.

Next is the Analysing Performance stage, which focuses on preprocessing the extracted text by organising it into sentences, manually labelling the data for sentiment analysis, ESG classification, and FLS classification, and testing the FinBERT algorithm on this labelled dataset. Once testing is complete, the FinBERT model is fine-tuned using the labelled corpus to enhance its performance and adaptability for the specific tasks of this study. This stage also addresses Objective 2, as it involves testing the FinBERT model on the prepared data and fine-tuning it to enhance its performance for the specific tasks of this study.

Finally, the Evaluation and Results stage assesses the baseline and fine-tuned model's performance and provides insights into the effectiveness of the model in achieving this study's objectives. This stage directly addresses Objective 3 as it provides an analysis of the model's ability to conduct sentiment analysis, ESG evaluation, and forward-looking statement classification on financial texts, thus presenting insights into its effectiveness in the context of Malaysian financial institutions.

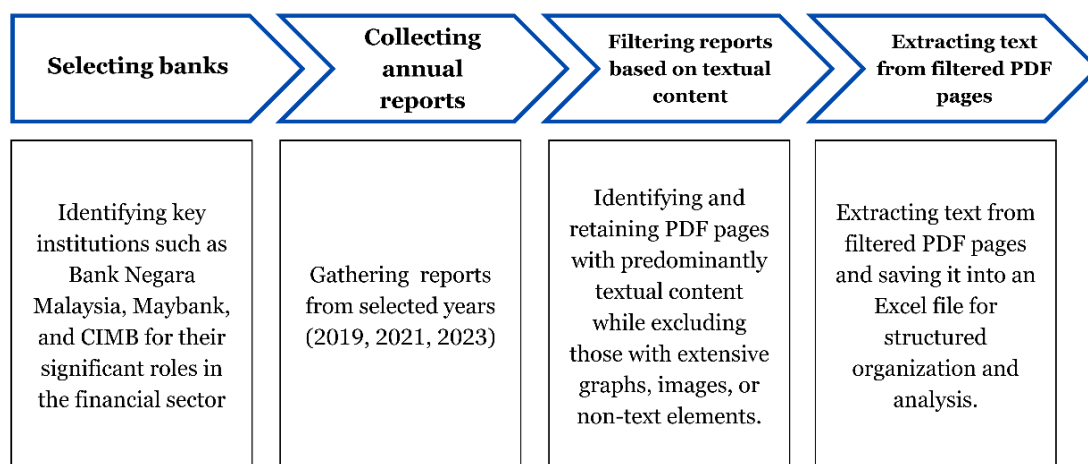
### 3.2.1 Literature Review

The study begins with a thorough literature review exploring the intersection between AI, NLP, and finance, with a focus on the Malaysian financial industry. It covers five key areas: AI's transformative role in sentiment analysis, ESG classification, and FLS classification; advancements in NLP and LLMs like BERT and GPT, and their applications in finance; the link between financial literacy and practice, with a focus on LLMs as tools to democratise financial knowledge; comparisons of architecture and application of domain-specific NLP models like FinBERT and BloombergGPT; and Malaysia's evolving financial landscape, emphasising digitalisation efforts, the role of Bank Negara Malaysia, and FinBERT's potential role in the local economy. Together, these sections provide a solid foundation for understanding this study's focus on AI-driven financial innovations.

### 3.2.2 Data Collection

The data collection phase is one of the most critical stages as it lays the groundwork for subsequent analysis.

**Figure 6.**  
*Flowchart of the Data Collection Process*



The phase begins with involves identifying key financial institutions that play a significant role in the financial sector. For this study, the chosen institutions are Bank Negara Malaysia, as the central bank, and two prominent local banks, Maybank and CIMB.

Next, annual reports from the selected institutions are collected for specific years (2019, 2021 and 2023). These reports, typically published as PDFs, serve as the primary data source for the study, providing comprehensive information on financial performance, strategies, and market trends.

Once the reports are collected, a thorough review is conducted to identify pages that are mostly textual. Pages from relevant sections of the reports that primarily contain paragraphs of text are kept, while those with extensive graphs, images, or other non-text elements are excluded. The filtering process ensures that the data focuses on textual content that is compatible with the FinBERT model. Additionally, it helps curate a dataset that is optimised for tasks like sentiment analysis, ESG evaluation, and identifying forward-looking statements.

**Figure 7.**  
Sample of Annual Report PDF



After the textual pages are identified, the text is extracted and then organised and saved into an Excel spreadsheet for further processing. The Python library that will be utilised is PyPDF, which enables text extraction from PDF files using the 'PdfReader' class to access and process individual pages. Below is a table detailing the codes and their purposes:

**Table 8.**  
Python Library Code and Their Purpose

| Code               | Purpose                                             |
|--------------------|-----------------------------------------------------|
| !pip install pypdf | Installs the PyPDF library for PDF text extraction. |

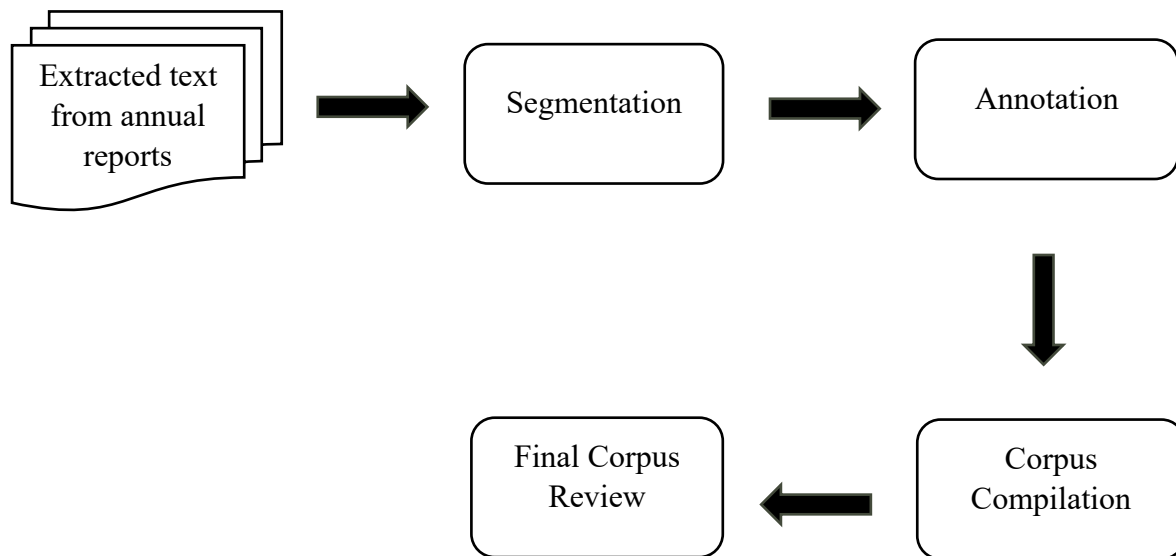
|                                                 |                                                      |
|-------------------------------------------------|------------------------------------------------------|
| <code>from pypdf import PdfReader</code>        | Imports the PdfReader class from the PyPDF library.  |
| <code>reader = PdfReader(file.pdf)</code>       | Loads the PDF file for processing.                   |
| <code>page = reader.pages[0]</code>             | Accesses a specific page (i.e., the first page).     |
| <code>print(page.extract_text())</code>         | Extracts and prints the text from the selected page. |
| <code>for i in range(len(reader.pages)):</code> | Iterates through all pages in the PDF.               |
| <code>page = reader.pages[i];</code>            | Extracts and prints text from each page in the loop. |
| <code>print(page.extract_text())</code>         |                                                      |

The next phase, text preprocessing, will refine the dataset to make it ready for use with the FinBERT model, enabling the analysis of sentiment, ESG evaluation, and the identification of forward-looking statements.

### 3.2.3 Text Preprocessing

The text preprocessing phase builds on the data collection phase by refining the extracted textual data for analysis. This phase ensures that the data is structured, labelled, and organised in a manner suitable to be input into the FinBERT model in the following phase.

**Figure 8.**  
Flowchart of Text Preprocessing Process



The first step involves segmenting the extracted text, initially in paragraph form, into individual sentences. Segmentation allows for a more granular level of analysis, ensuring precise labelling and classification. Each sentence is stored in a structured format within the Excel sheet, where each row represents a unique sentence.

**Figure 9.**  
Sample of Segmented Sentences

|    | A                                                                                                                                                                                                |
|----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1  | <b>Original Sentence</b>                                                                                                                                                                         |
| 2  | As a highly open economy, the Malaysian economy was affected by these global developments.                                                                                                       |
| 3  | weakness in investment activity and supply disruptions in the commodities sector also affected domestic economic activity.                                                                       |
| 4  | Economic growth thus expanded by 4.3% in 2019 (2018: 4.7%), driven by private sector spending.                                                                                                   |
| 5  | In particular, household spending was supported by continued income and employment growth, as well as Government measures.                                                                       |
| 6  | Meanwhile, price pressures were subdued throughout the year.                                                                                                                                     |
| 7  | The low inflation was due mainly to policies on retail fuel prices and festive price controls, as well as the changes in consumption tax policy in 2018, which resulted in lower price pressures |
| 8  | The decision to lower the OPR was made in view of providing a conducive monetary environment for a steady growth path amid price stability.                                                      |
| 9  | In November 2019, the SRR ratio was reduced from 3.50% to 3.00% to continue supporting the efficient functioning of the domestic financial markets.                                              |
| 10 | Malaysia maintains a flexible exchange rate policy for ringgit.                                                                                                                                  |

Once the sentences are segmented, manual annotation is conducted by an industry expert. Each sentence is carefully reviewed and labelled based on three key criteria: For sentiment analysis, sentences are tagged with *Positive*, *Negative* or *Neutral* based on their sentiment scores (continuous numeric values from +1 for positive to -1 for negative); For ESG classification, sentences are categorised into *Environmental*, *Social*, or *Governance*. Sentences

that do not align with these labels are marked as *None*; For FLS, sentences are marked with *Specific-FLS* (if they discuss future plans or projections specific to the company), *Non-specific FLS* (if they are future-oriented but contain general statements that could apply to any company, such as cautionary language or risk disclosures), or *Not-FLS* (if they do not reference the future). After annotation, the labelled sentences are compiled into a structured corpus, with each task organized into its own sheet within the Excel file. Each sheet corresponds to a specific task: one for Sentiment Analysis (SA), another for ESG Classification (ESG), and a third for Forward-Looking Statements (FLS). Within each sheet, sentences are given unique identifiers based on the task. For example, sentences in the sentiment analysis sheet are labelled with IDs such as *SA01*, *SA02*, and so on. Similarly, sentences for ESG and FLS classification use *ESG01*, *ESG02*, and *FLS01*, *FLS02* respectively.

**Figure 10.**  
*Sample of Excel Corpus Structure for Sentiment Analysis*

|    | A   | B                                                                                                                                                                                                | C         |
|----|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| 1  | ID  | Original Sentence                                                                                                                                                                                | Sentiment |
| 2  | SA1 | Globally, there were unresolved trade tensions, a slowdown in investments and trade activity, heightened financial market volatility, country-specific risks and geopolitical uncertainties.     | Negative  |
| 3  | SA2 | As a highly open economy, the Malaysian economy was affected by these global developments.                                                                                                       | Negative  |
| 4  | SA3 | weakness in investment activity and supply disruptions in the commodities sector also affected domestic economic activity.                                                                       | Negative  |
| 5  | SA4 | Economic growth thus expanded by 4.3% in 2019 (2018: 4.7%), driven by private sector spending.                                                                                                   | Positive  |
| 6  | SA5 | In particular, household spending was supported by continued income and employment growth, as well as Government measures.                                                                       | Positive  |
| 7  | SA6 | Meanwhile, price pressures were subdued throughout the year.                                                                                                                                     | Neutral   |
| 8  | SA7 | The low inflation was due mainly to policies on retail fuel prices and festive price controls, as well as the changes in consumption tax policy in 2018, which resulted in lower price pressures | Neutral   |
| 9  | SA8 | The decision to lower the OPR was made in view of providing a conducive monetary environment for a steady growth path amid price stability.                                                      | Positive  |
| 10 | SA9 | In November 2019, the SRR ratio was reduced from 3.50% to 3.00% to continue supporting the efficient functioning of the domestic financial markets.                                              | Positive  |

**Figure 11.**  
*Sample of Excel Corpus Structure for ESG Classification*

|    | A     | B                                                                                                                                                                                                                                                              | C            |
|----|-------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------|
| 1  | ID    | Original Sentence                                                                                                                                                                                                                                              | ESG Category |
| 2  | ESG1  | In particular, household spending was supported by continued income and employment growth, as well as Government measures.                                                                                                                                     | Social       |
| 3  | ESG2  | The low inflation was due mainly to policies on retail fuel prices and festive price controls, as well as the changes in consumption tax policy in 2018, which resulted in lower price pressures                                                               | Governance   |
| 4  | ESG3  | Our staff facilitate this process by presenting their surveillance analysis, simulation results, research findings and forecasts to the members                                                                                                                | Governance   |
| 5  | ESG4  | At each meeting, the MPC decides on whether the current level of the Overnight Policy Rate (OPR) – the indicator of the monetary policy stance – is at an appropriate level to safeguard price stability while supporting economic activity                    | Governance   |
| 6  | ESG5  | One of our roles is to maintain the efficient and effective functioning of the foreign exchange market.                                                                                                                                                        | Governance   |
| 7  | ESG6  | In this regard, we continue to ensure that volatility in the exchange rate is not excessive so as not to disrupt domestic economic activity such as trade and investments.                                                                                     | Governance   |
| 8  | ESG7  | We also conducted research on longer-term and structural issues confronting the economy.                                                                                                                                                                       | Governance   |
| 9  | ESG8  | We also provided advisory support to the Government on reviewing national investment policy through participation in key working groups such as for the New Industrial Masterplan and for the Twelfth Malaysia Plan (Rancangan Malaysia Kedua Belas, RMKe-12). | Governance   |
| 10 | ESG9  | Research emphasis during the year was on ways to reinvigorate private investments in the country                                                                                                                                                               | Governance   |
| 11 | ESG10 | As a financial advisor to the Government, we are a member of the Fiscal Policy Committee (FPC) and Debt Management Committee (DMC), providing input on strategies to strengthen the country's fiscal framework.                                                | Governance   |

**Figure 12.**  
*Sample of Excel Corpus Structure for Forward Looking Statement*

|    | A     | B                                                                                                                                                                                                                                           | C                  |
|----|-------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------|
| 1  | ID    | Original Sentence                                                                                                                                                                                                                           | FLS Classification |
| 2  | FLS1  | At each meeting, the MPC decides on whether the current level of the Overnight Policy Rate (OPR) – the indicator of the monetary policy stance – is at an appropriate level to safeguard price stability while supporting economic activity | Not-FLS            |
| 3  | FLS2  | As the effect of monetary policy takes at least two or three quarters to be felt in the economy, the MPC, in making its decisions, assesses and deliberates on the outlook for both domestic economic growth and inflation                  | Non-specific FLS   |
| 4  | FLS3  | The MPC also takes into consideration the risk that financial imbalances may pose to the broader economy.                                                                                                                                   | Not-FLS            |
| 5  | FLS4  | Malaysia maintains a flexible exchange rate policy for ringgit.                                                                                                                                                                             | Not-FLS            |
| 6  | FLS5  | This is achieved by providing sufficient liquidity at all times in the domestic foreign exchange market, including through our two-way foreign exchange interventions.                                                                      | Not-FLS            |
| 7  | FLS6  | Work focused on three key areas of reform including establishing national investment objectives, streamlining investment promotion and investor services as well as improving the efficiency of tax incentive administration.               | Not-FLS            |
| 8  | FLS7  | Research emphasis during the year was on ways to reinvigorate private investments in the country                                                                                                                                            | Not-FLS            |
| 9  | FLS8  | As a financial advisor to the Government, we are a member of the Fiscal Policy Committee (FPC) and Debt Management Committee (DMC), providing input on strategies to strengthen the country's fiscal framework.                             | Not-FLS            |
| 10 | FLS9  | With the world facing the outbreak of COVID-19, the Bank will focus on managing the impact of this unprecedented global health crisis on the Malaysian economy.                                                                             | Specific-FLS       |
| 11 | FLS10 | In this regard, the MPC will continue to assess all relevant data as well as economic and monetary developments in arriving at the appropriate monetary policy response.                                                                    | Non-specific FLS   |

Once the corpus is fully annotated and organised, a final review is conducted. This is to ensure the quality and consistency of the labelled data. The annotations are cross-checked to identify and resolve any inconsistencies or errors in labelling and the structure of the corpus is verified to confirm that each sheet is correctly formatted and aligned with the requirements of the analysis tasks.

The final reviewed dataset is named the Malaysian Financial Text Corpus (MFTC). This corpus serves as a comprehensive and reliable resource for sentiment analysis, ESG classification, and forward-looking statement identification in the context of Malaysian financial institutions.

### 3.2.4 Testing FinBERT Algorithm

The testing phase involves evaluating FinBERT's performance on the annotated sentences from the Malaysian Financial Text Corpus (MFTC). This involves loading the labelled sentences, feeding them into the model for predictions, and comparing the results with manual annotations. Three distinct FinBERT models, from Yang's FinBERT-20, are used for sentiment analysis, ESG classification, and FLS classification.

Before testing, a few necessary libraries are installed. The primary library used is Huggingface Transformers, which enables the usage of the pre-trained FinBERT models. The

labelled sentences from the MFTC are stored in an Excel file, with separate sheets for each task and the pandas library is used to load the data. Below is a table of the relevant libraries and their purpose:

**Table 9.**  
*Library Overview*

| Library                       | Code Snippet                                      | Purpose                                                                            |
|-------------------------------|---------------------------------------------------|------------------------------------------------------------------------------------|
| transformer                   | <code>!pip install transformers</code>            | To load and use pre-trained FinBERT models                                         |
| pandas                        | <code>!pip install pandas</code>                  | Manage and process Excel data                                                      |
| BertTokenizer                 | <code>import BertTokenizer</code>                 | Used to convert text into tokenized format                                         |
| BertForSequenceClassification | <code>import BertForSequenceClassification</code> | Pre-trained BERT model                                                             |
| pipeline                      | <code>import pipeline</code>                      | Utility function that simplifies the process of using pre-trained models for tasks |

Next, depending on the task that is being evaluated, the specific fine-tuned FinBERT model is loaded. The table below details the specifics of each model.

**Table 10.**  
*Code Snippet of Fine-tuned FinBERT Models*

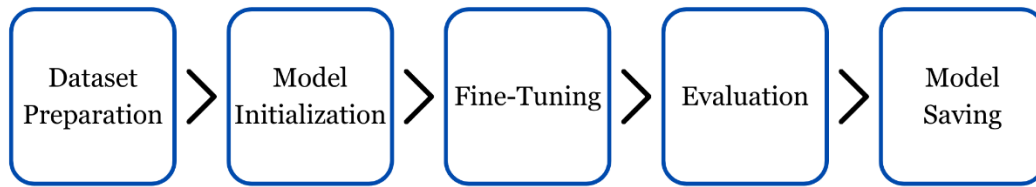
| Task               | Code Snippet                                                                                                                                                                                                                                                 |
|--------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Sentiment Analysis | <pre>finbert = BertForSequenceClassification.from_pretrained('yiyanghkust/finbert-tone', num_labels=3) tokenizer = BertTokenizer.from_pretrained('yiyanghkust/finbert-tone')  nlp = pipeline("sentiment-analysis", model=finbert, tokenizer=tokenizer)</pre> |
| ESG Classification | <pre>finbert = BertForSequenceClassification.from_pretrained('yiyanghkust/finbert-esg', num_labels=4) tokenizer = BertTokenizer.from_pretrained('yiyanghkust/finbert-esg')  nlp = pipeline("text-classification", model=finbert, tokenizer=tokenizer)</pre>  |
| FLS Classification | <pre>finbert = BertForSequenceClassification.from_pretrained('yiyanghkust/finbert-fls', num_labels=3) tokenizer = BertTokenizer.from_pretrained('yiyanghkust/finbert-fls')  nlp = pipeline("text-classification", model=finbert, tokenizer=tokenizer)</pre>  |

Once all necessary libraries are installed and the labelled sentences are loaded, each sentence is processed individually using the pipeline. The predicted labels are then appended to the dataset, creating a structured output that combines the original sentences with their corresponding predictions. Finally, the updated dataset is saved for further analysis, ensuring the results are systematically recorded and ready for evaluation.

### 3.2.5 Fine-Tuning the FinBERT Model

In this phase, the pretrained FinBERT model is fine-tuned using a subset of the labelled data from the MFTC. This process involves adapting the model to better understand the nuances of financial texts specific to the Malaysian context, enhancing its performance on downstream tasks like sentiment analysis and ESG classification.

**Figure 13.**  
*Overview of Fine-tuning Process*



In the Dataset preparation phase, a portion of data from the MTFC is set aside before the FinBERT model is tested to be used as the dataset for the fine-tuning process. The dataset is then split into training, validation and test subsets, ensuring an appropriate distribution of data across these splits to facilitate model training and evaluation.

For the model initialisation, the pretrained FinBERT model from Hugging Face's model hub is used as the starting point. This model, trained on large-scale financial text data, serves as the base for further fine-tuning. The dataset is tokenized using the FinBERT tokenizer, which ensures that the input format aligns with the model's requirements. Each sentence in the dataset is tokenized with truncation and padding applied to a maximum length of 128 tokens, making it compatible with the model's architecture.

During the fine-tuning phase, hyperparameters and training arguments are defined to optimise the model's performance. The training setup includes a learning rate of  $2 \times 10^{-5}$ , a batch size of 32, weight decay of 0.01 for regularisation, and five training epochs. The Hugging Face Trainer class is employed to streamline the training process, handling tasks such as batching, optimization, and validation. The model's performance is evaluated at the end of each epoch, and the checkpoint with the best accuracy on the validation set is saved for future use.

Once the fine-tuning process is complete, the model is evaluated on the test set to assess its generalisation capability. This evaluation provides insights into how well the model performs on unseen data. After the evaluation, the fine-tuned model is saved for future use.

### 3.2.6 Evaluation and Results

For this study, the performance metrics that are used to evaluate the pretrained and fine-tuned FinBERT models are accuracy and a confusion matrix. These metrics provide a comprehensive assessment of how well the models perform on tasks such as sentiment analysis, ESG classification, and FLS classification in a context-specific situation.

Accuracy is the primary metric used to evaluate the overall performance of the models. To calculate accuracy, a proportion of correct predictions (both true positives and true negatives) out of the total number of predictions is measured. The formula for accuracy is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP (True Positive): The number of correctly predicted positive instances.
- TN (True Negative): The number of correctly predicted negative instances.
- FP (False Positive): The number of instances that were incorrectly predicted as positive
- FN (False Negative): The number of instances that were incorrectly predicted as negative.

In addition to accuracy, a confusion matrix will also be utilised to better understand the types of errors made by the model. The confusion matrix provides a breakdown of the model's predictions across all classes and highlights the number of true positives, true negatives, false positives, and false negatives for each class. For multi-class classification tasks, such as sentiment analysis, ESG classification, and FLS classification, the confusion matrix extends to show the relationships between actual and predicted labels for each class.

From the confusion matrix, additional metrics such as precision, recall and F1-score can be derived:

- Precision measures the accuracy of positive predictions of each class:

$$Precision = \frac{TP}{TP + FP}$$

- Recall measures the ability of the model to correctly identify positive instances for each class:

$$Recall = \frac{TP}{TP + FN}$$

- F1-score provides a balanced measure of precision and recall, especially useful when dealing with imbalanced classes:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

These performance metrics serve as the foundation for evaluating the FinBERT models in this study. They provide both a high-level overview of model performance and a detailed breakdown of how the models perform on specific tasks, helping to identify areas where improvements can be made, especially during the fine-tuning phase.

### 3.3 Chapter Summary

In this chapter, the methodology of the study is outlined summarising the approach taken across six distinct research phases. It provides a detailed explanation of how the project is carried out, from the data collection and preprocessing to testing and fine-tuning the model. The implementation and testing of the study will be presented in the following chapter.

## **Chapter 4: Testing, Fine-Tuning of the FinBERT Model**

### **4.1 Introduction**

This chapter presents the testing phase of the FinBERT model using annotated sentences from the Malaysian Financial Text Corpus (MFTC), previously introduced in Chapter 3. The objective of this phase is to evaluate FinBERT’s performance across three core tasks: Sentiment Analysis, ESG Classification, and Forward-Looking Statement (FLS) Classification, which represent critical dimensions in understanding financial disclosures, particularly within the Malaysian banking sector.

In this study, the FinBERT model is tested by running a test set of unlabelled sentences, and the model’s predictions are benchmarked against gold-standard annotations by an industry expert. The model’s performance is measured using standard classification metrics: accuracy and a confusion matrix, from which precision, recall and F1-score can be derived. Beyond quantitative methods, a qualitative error analysis is also conducted to understand the nature of misclassifications. In particular, factors such as the complexity of annual report structures – where a mix of text, tables, and images may disrupt text extraction – may negatively impact classification accuracy. The insights gained from this phase guide the fine-tuning process, helping to refine FinBERT’s ability to handle financial text from the Malaysian banking sector.

### **4.2 Dataset Composition**

#### **4.2.1 Structure of Dataset**

The Malaysian Financial Text Corpus, previously introduced in Chapter 3, comprises of 3,067 sentences extracted from a total of nine annual reports. These reports include documents from Bank Negara Malaysia (BNM), Malayan Banking Berhad (Maybank), and Commerce International Merchant Bankers Berhad (CIMB), covering the years 2019, 2021, and 2023. The sentences were sourced primarily from the narrative sections of the reports, excluding tables,

figures, appendices, and any non-textual elements to ensure the content is compatible with the model's requirements.

The dataset was split into three subsets for model training, testing and evaluation respectively. Specifically, 20% of the dataset (613 sentences) was allocated for testing task-specific FinBERT models that have already been fine-tuned on large-scale financial text corpora; 70% (2,147 sentences) was reserved for fine-tuning a context-specific model for the Malaysian financial industry; and 10% (307 sentences) was kept aside as a validation set to be used during fine-tuning to monitor performance and to prevent over-fitting. This distribution ensures sufficient data for the model training while maintain an unbiased test set for performance evaluation.

**Table 11.**  
*Distribution of Sentences by Report*

| Institution                                                 | 2019         | 2021       | 2023       | <b>Total</b> |
|-------------------------------------------------------------|--------------|------------|------------|--------------|
| Bank Negara Malaysia<br>(BNM)                               | 571          | 403        | 296        | <b>1,270</b> |
| Malayan Banking Berhad<br>(Maybank)                         | 289          | 284        | 306        | <b>879</b>   |
| Commerce International<br>Merchant Bankers Berhad<br>(CIMB) | 296          | 306        | 316        | <b>918</b>   |
| <b>Total</b>                                                | <b>1,156</b> | <b>993</b> | <b>918</b> | <b>3,067</b> |

The table above represents the number of sentences extracted from each annual report included in the dataset. The data highlights the relative volume of content per report, providing a quantitative overview of the sentence distribution.

**Figure 14.**  
*Percentage of Sentences by Institution*

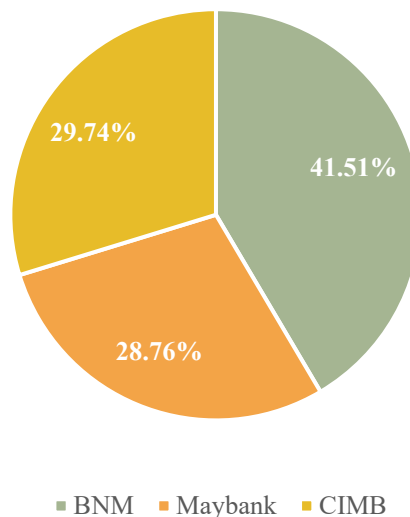


Figure 14 illustrates the percentage distribution of sentences across the selected reports. The chart visualises. Each segment represents a report's contribution to the total sentence count. From this, a clear comparison of their relative proportions can be made, highlighting which reports have a greater or lesser presence within the overall dataset.

A separate metadata file is maintained to preserve the original source and year of each sentence. This allows for transparency and traceability despite the random shuffling, ensuring that the origin of any sentence can be referenced during analysis.

**Figure 15.**  
*Structure of Metadata*

| ID   | Sentences                                                                                                                                                                                                     | Source | Year | Sentiment Analysis | ESG Classification | FLS Classification |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|------|--------------------|--------------------|--------------------|
| BNM1 | A key role of the Bank is to implement a monetary policy that maintains a low and predictable pace of increase in the prices of goods and services, taking into consideration existing economic developments. | BNM    | 2019 | Neutral            | None               | Not-FLS            |
| BNM2 | Globally, there were unresolved trade tensions, a slowdown in investments and trade activity, heightened financial market volatility, country-specific risks and geopolitical uncertainties.                  | BNM    | 2019 | Negative           | None               | Not-FLS            |
| BNM3 | As a highly open economy, the Malaysian economy was affected by these global developments.                                                                                                                    | BNM    | 2019 | Negative           | None               | Not-FLS            |
| BNM4 | In addition, weakness in investment activity and supply disruptions in the commodities sector also affected domestic economic activity.                                                                       | BNM    | 2019 | Negative           | None               | Not-FLS            |
| BNM5 | Economic growth thus expanded by 4.3% in 2019 (2018: 4.7%), driven by private sector spending.                                                                                                                | BNM    | 2019 | Neutral            | None               | Not-FLS            |
| BNM6 | In particular, household spending was supported by continued income and employment growth, as well as Government measures.                                                                                    | BNM    | 2019 | Positive           | None               | Not-FLS            |
| BNM7 | Headline inflation averaged at 0.7% in 2019 (2018: 1.0%).                                                                                                                                                     | BNM    | 2019 | Positive           | None               | Not-FLS            |
| BNM8 | The low inflation was due mainly to policies on retail fuel prices and festive price controls, as well as the changes in consumption tax policy in 2018, which resulted in lower price pressures.             | BNM    | 2019 | Positive           | Governance         | Not-FLS            |

Figure 15 shows the structure of the metadata file. The file consists of five key attributes: a unique ID for each bank (i.e. BNM1, BNM2, BNM3 for Bank Negara Malaysia; MB1, MB2, MB3 for Maybank and CIMB1, CIMB2, CIMB3 for CIMB), the extracted sentence from the corresponding report, the source and year from which the related sentence originates from, and the gold-standard annotations provided by the industry expert. This structure ensures ease of traceability while establishing a foundation for evaluating the model's performance against expert-labelled data.

The annotations consists of the three classifications tasks evaluated in this study: sentiment analysis (positive, neutral, negative), ESG classification (environmental, social, governance, or none) and forward-looking statement classification (specific-FLS, non-specific-FLS, not-FLS). The table below illustrates the label distribution for each task.

**Table 12.**  
*Distribution of Gold Standard Annotations*

| Task               | Label            | Count | Percentage (%) |
|--------------------|------------------|-------|----------------|
| Sentiment Analysis | Positive         | 1,363 | 44.44%         |
|                    | Neutral          | 1,390 | 45.32%         |
|                    | Negative         | 314   | 10.24%         |
| ESG Classification | Environmental    | 588   | 19.17%         |
|                    | Social           | 328   | 10.69%         |
|                    | Governance       | 1414  | 46.10%         |
|                    | None             | 737   | 24.03%         |
| FLS Classification | Specific-FLS     | 130   | 4.24%          |
|                    | Non-specific FLS | 289   | 9.42%          |
|                    | Not-FLS          | 2,648 | 86.34%         |

Table 12 shows the distribution of labels assigned during the annotation process. For each label, the number of sentences and their corresponding percentage relative to the overall dataset is noted. This provides an overview into the balance of the dataset while also aiding in identifying any potential class imbalances that may influence model training and evaluation. The distribution provides a look into the overall thematic and tonal compositions of financial reports from Malaysian institutions.

#### **4.2.2 Statistical Analysis of Dataset**

The statistical analysis of the dataset provides a look into the linguistic characteristics and patterns within the Malaysian Financial Text Corpus (MFTC) used for evaluating the FinBERT model. The analysis includes word frequency counts and a breakdown of the most commonly used words by year to identify notable themes and vocabulary patterns across reports.

##### **4.2.2.1 Word Frequency Analysis**

To gain an initial overview of the lexical content in the dataset, a word frequency analysis was conducted on all 3,067 sentences. Prior to conducting the word analysis on the dataset, several preprocessing steps were applied to ensure data consistency and enhance the quality of the output. These steps included converting all text to lowercase, removing punctuation and possessive forms, eliminating common stopwords, and filtering out frequently recurring non-informative words such as “The Bank,” “The Group,” and “Malaysia.” etc. using a custom stopword list implemented with the NLTK library. Words with fewer than three characters were also excluded and lemmatization was applied to reduce words to their base forms, thereby improving the accuracy of the term frequently analysis. A word cloud was subsequently generated to visualise the most frequent terms across the entire dataset, and the top twenty terms were listed to identify key themes present in the text.

**Figure 16.**

*Code Snippet of Word Frequency Analysis and Word Cloud Generation*

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
import nltk
import re
import pandas as pd
from nltk.stem import WordNetLemmatizer
from nltk.corpus import wordnet
from collections import Counter

#nltk resources
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('omw-1.4')

#load dataset
df = pd.read_excel("MTFC Dataset.xlsx", sheet_name="Combined")
sentences = df['Sentences'].tolist()

#combine text
text = " ".join(sentences).lower()

#remove custom words
custom_phrases = ["the bank", "malaysia", "the group", "year", "programme",
"bank", "business", "continued", "continue", "2021", "2019", "2023",
"well", "across", "financial"]

for phrase in custom_phrases:
text = text.replace(phrase, "")

#remove possessive 's
text = re.sub(r"'s\b", "", text)

#tokenise
words = re.findall(r'\b\w+\b', text)

#define stopwords
stop_words = set(stopwords.words('english'))
```

```

extra_stops = {'s', 'may', 'shall', 'will', 'also', 'can', 'could',
'would', 'ing', 'nt', 'us',}
stop_words.update(extra_stops)

#filter
filtered_words = [word for word in words if word not in stop_words and word
not in custom_phrases and len(word) > 2]

#lemmatization
lemmatizer = WordNetLemmatizer()
lemmatized_words = [lemmatizer.lemmatize(word) for word in filtered_words]

#join words
filtered_text = " ".join(filtered_words)
filtered_text = " ".join(lemmatized_words)

#count word frequencies
word_freq = Counter(filtered_words)

#top 20 most common words
top_terms = word_freq.most_common(20)

word_freq = Counter(lemmatized_words)
top_terms = word_freq.most_common(20)

#print the top terms
print("Top 20 most frequent terms:")
for word, freq in top_terms:
print(f"{word}: {freq}")

#word cloud
wordcloud = WordCloud(width=800, height=400, background_color='white',
relative_scaling=1.0).generate(filtered_text)
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()

```

**Figure 17.**  
*Word Cloud of Most Frequently Used Terms in the Dataset*



The word cloud above visualises the most frequently occurring terms in the MTFC dataset, which consist of financial and policy-related documents. It was observed that terms such as “risk”, “financing”, “sustainability”, “support” and “policy” were mentioned more frequently indicating a strong emphasis on financial risk management, sustainable development, and institutional support mechanisms in the reports.

**Table 13.**  
*Top 20 Most Frequent Terms*

| Term        | Frequency |
|-------------|-----------|
| risk        | 497       |
| policy      | 238       |
| climate     | 229       |
| sector      | 208       |
| management  | 206       |
| financing   | 201       |
| market      | 189       |
| growth      | 169       |
| customer    | 168       |
| sustainable | 166       |
| economy     | 163       |

|                |     |
|----------------|-----|
| support        | 162 |
| service        | 160 |
| sustainability | 153 |
| economic       | 150 |
| billion        | 150 |
| institution    | 147 |
| global         | 146 |
| digital        | 142 |
| industry       | 137 |

The table above depicts the top twenty most frequently mentioned terms from the dataset. The frequencies reflect the number of times each term appears and provides a clearer view on the dominant themes such as risk considerations, climate-related issues, financial management, and sustainability efforts.

#### **4.2.3 Test Set Composition**

The testing dataset for the FinBERT model consists of 613 unlabelled sentences randomly sources from five annual reports: BNM 2019, BNM 2021, BNM 2023, Maybank 2019, and CIMB 2019. These sentences, which make up 20% of the overall dataset, were used to evaluate the model’s ability to perform sentiment analysis, ESG classification, and forward-looking statement classification.

**Figure 18.**  
*Structure of Test Set*

| C  |      | D                                                                                                                                                                                                                                                                                                                                                                                    |  |
|----|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|
| 1  | ID   | Sentences                                                                                                                                                                                                                                                                                                                                                                            |  |
| 2  | SA1  | Under the initiative, Maybank Islamic created opportunities for social enterprises to engage with Maybank employees and the public and promote awareness about the role social enterprises have to play in sustainable social development.                                                                                                                                           |  |
| 3  | SA2  | Work together with other banks, clients, governments and policymakers and support each other in developing capabilities and the necessary methodologies to measure climate impact, align with global and local climate goals, and help transition to low-carbon, climate-resilient economy.                                                                                          |  |
| 4  | SA3  | Our advocacy aimed to ensure the CMIM continues to reflect the needs of the region.                                                                                                                                                                                                                                                                                                  |  |
| 5  | SA4  | This facilitates effective monetary policy transmission as adjustments to loan repayments will take effect following an OPR change.                                                                                                                                                                                                                                                  |  |
| 6  | SA5  | A prospective borrower could also be offered suggestions to improve his or her financial position, such as ways to generate more income or rationalise existing debt and other commitments before reapplying.                                                                                                                                                                        |  |
| 7  | SA6  | While this is a positive development and a sign of progress in the country's health outcomes, this development naturally contributes to higher medical costs.                                                                                                                                                                                                                        |  |
| 8  | SA7  | Key Payment Trends Usage of e-payment services continued to gain further traction.                                                                                                                                                                                                                                                                                                   |  |
| 9  | SA8  | Recommendations included a collaboration platform between authorities, industry, academic and civil society in designing education policies; strengthening interlinkages between social security benefits and Active Labour Market Policies (ALMPs); as well as extension of employment insurance and training scheme coverage to vulnerable groups.                                 |  |
| 10 | SA9  | We have also partnered with Business for Social Responsibility (BSR) to strengthen sustainable business.                                                                                                                                                                                                                                                                             |  |
| 11 | SA10 | We continue to work closely with our international counterparts in support of efforts to promote global financial stability.                                                                                                                                                                                                                                                         |  |
| 12 | SA11 | We also continue to intensify our initiative to bridge the knowledge gap in the area of fintech at all levels of the organisation.                                                                                                                                                                                                                                                   |  |
| 13 | SA12 | In addition to formal structures and processes, the Bank emphasises a strong risk culture that promotes risk awareness and open communication on risk.                                                                                                                                                                                                                               |  |
| 14 | SA13 | Through strong business performance and ongoing growth, we continue to distribute value to our customers, communities and employees.                                                                                                                                                                                                                                                 |  |
| 15 | SA14 | At the same time, we have kept our customers abreast with the latest local and global market Maybank's Response to COVID-19 developments through regular e-updates and webinars for our affluent and high net worth customers while Maybank Kim Eng Research continues to provide daily, weekly and quarterly market and strategy updates as well as recommendations to its clients. |  |
| 16 | SA15 | To achieve this goal, the MPC decides on the Overnight Policy Rate (OPR) based on its assessment of domestic inflation and economic growth prospects.                                                                                                                                                                                                                                |  |
| 17 | SA16 | During the year, we reviewed 34 cases of breaches of our laws and regulations, including those committed by individuals and entities not regulated by the Bank.                                                                                                                                                                                                                      |  |

The figure above shows the composition of the test set across the selected reports. To ensure an even distribution of content from different sources, all sentences within the test set were randomly shuffled, and the first 613 sentences were selected for evaluation. This step eliminates the possibility of an ordering bias that that may be a result of the structure or flow of the original reports in addition to allowing the model to evaluate each sentence independently.

Each sentence in the test set is paired with unique identifiers corresponding to their respective tasks, such as SA1, SA2 for Sentiment Analysis, or ESG1, ESG2 for ESG Classification, etc. These identifiers are used to track and refer to specific sentences throughout the evaluation process, ensuring that the process is organised and specific data points are be clearly attributed to their result.

Figures 17, 18 and 19 show the structure of the test sets used for evaluating the FinBERT model. All three test sets are derived from the same pool of 613 sentences selected

from the Malaysian Financial Text Corpus. Each test set shares a common format comprising of two columns: a unique task-specific ID (“ID”) and the corresponding sentence (“Sentences”).

**Figure 19.**  
*Test Set Structure for Sentiment Analysis*

| ID   | Sentences                                                                                                                                                                                                                                                                                                                                            |
|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| SA1  | Under the initiative, Maybank Islamic created opportunities for social enterprises to engage with Maybank employees and the public and promote awareness about the role social enterprises have to play in sustainable social development.                                                                                                           |
| SA2  | Work together with other banks, clients, governments and policymakers and support each other in developing capabilities and the necessary methodologies to measure climate impact, align with global and local climate goals, and help transition to low-carbon, climate-resilient economy.                                                          |
| SA3  | Our advocacy aimed to ensure the CMIM continues to reflect the needs of the region.                                                                                                                                                                                                                                                                  |
| SA4  | This facilitates effective monetary policy transmission as adjustments to loan repayments will take effect following an OPR change.                                                                                                                                                                                                                  |
| SA5  | A prospective borrower could also be offered suggestions to improve his or her financial position, such as ways to generate more income or rationalise existing debt and other commitments before reapplying.                                                                                                                                        |
| SA6  | While this is a positive development and a sign of progress in the country's health outcomes, this development naturally contributes to higher medical costs.                                                                                                                                                                                        |
| SA7  | Key Payment Trends Usage of e-payment services continued to gain further traction.                                                                                                                                                                                                                                                                   |
| SA8  | Recommendations included a collaboration platform between authorities, industry, academic and civil society in designing education policies; strengthening interlinkages between social security benefits and Active Labour Market Policies (ALMPs); as well as extension of employment insurance and training scheme coverage to vulnerable groups. |
| SA9  | We have also partnered with Business for Social Responsibility (BSR) to strengthen sustainable business.                                                                                                                                                                                                                                             |
| SA10 | We continue to work closely with our international counterparts in support of efforts to promote global financial stability.                                                                                                                                                                                                                         |

**Figure 20.**  
*Test Set Structure for ESG Classification*

| ID    | Sentences                                                                                                                                                                                                                                                                                                                                            |
|-------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| ESG1  | Under the initiative, Maybank Islamic created opportunities for social enterprises to engage with Maybank employees and the public and promote awareness about the role social enterprises have to play in sustainable social development.                                                                                                           |
| ESG2  | Work together with other banks, clients, governments and policymakers and support each other in developing capabilities and the necessary methodologies to measure climate impact, align with global and local climate goals, and help transition to low-carbon, climate-resilient economy.                                                          |
| ESG3  | Our advocacy aimed to ensure the CMIM continues to reflect the needs of the region.                                                                                                                                                                                                                                                                  |
| ESG4  | This facilitates effective monetary policy transmission as adjustments to loan repayments will take effect following an OPR change.                                                                                                                                                                                                                  |
| ESG5  | A prospective borrower could also be offered suggestions to improve his or her financial position, such as ways to generate more income or rationalise existing debt and other commitments before reapplying.                                                                                                                                        |
| ESG6  | While this is a positive development and a sign of progress in the country's health outcomes, this development naturally contributes to higher medical costs.                                                                                                                                                                                        |
| ESG7  | Key Payment Trends Usage of e-payment services continued to gain further traction.                                                                                                                                                                                                                                                                   |
| ESG8  | Recommendations included a collaboration platform between authorities, industry, academic and civil society in designing education policies; strengthening interlinkages between social security benefits and Active Labour Market Policies (ALMPs); as well as extension of employment insurance and training scheme coverage to vulnerable groups. |
| ESG9  | We have also partnered with Business for Social Responsibility (BSR) to strengthen sustainable business.                                                                                                                                                                                                                                             |
| ESG10 | We continue to work closely with our international counterparts in support of efforts to promote global financial stability.                                                                                                                                                                                                                         |

**Figure 21.**  
*Test Set Structure for FLS Classification*

| ID    | Sentences                                                                                                                                                                                                                                                                                                                                            |
|-------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| FLS1  | Under the initiative, Maybank Islamic created opportunities for social enterprises to engage with Maybank employees and the public and promote awareness about the role social enterprises have to play in sustainable social development.                                                                                                           |
| FLS2  | Work together with other banks, clients, governments and policymakers and support each other in developing capabilities and the necessary methodologies to measure climate impact, align with global and local climate goals, and help transition to low-carbon, climate-resilient economy.                                                          |
| FLS3  | Our advocacy aimed to ensure the CMIM continues to reflect the needs of the region.                                                                                                                                                                                                                                                                  |
| FLS4  | This facilitates effective monetary policy transmission as adjustments to loan repayments will take effect following an OPR change.                                                                                                                                                                                                                  |
| FLS5  | A prospective borrower could also be offered suggestions to improve his or her financial position, such as ways to generate more income or rationalise existing debt and other commitments before reapplying.                                                                                                                                        |
| FLS6  | While this is a positive development and a sign of progress in the country's health outcomes, this development naturally contributes to higher medical costs.                                                                                                                                                                                        |
| FLS7  | Key Payment Trends Usage of e-payment services continued to gain further traction.                                                                                                                                                                                                                                                                   |
| FLS8  | Recommendations included a collaboration platform between authorities, industry, academic and civil society in designing education policies; strengthening interlinkages between social security benefits and Active Labour Market Policies (ALMPs); as well as extension of employment insurance and training scheme coverage to vulnerable groups. |
| FLS9  | We have also partnered with Business for Social Responsibility (BSR) to strengthen sustainable business.                                                                                                                                                                                                                                             |
| FLS10 | We continue to work closely with our international counterparts in support of efforts to promote global financial stability.                                                                                                                                                                                                                         |

### **4.3 Experimental Setup**

Before deploying the FinBERT model, several preliminary steps were taken to setup the testing environment. This included preparing and formatting the dataset, selecting the appropriate pretrained FinBERT model from Yang’s FinBERT-20 according to the three tasks – sentiment analysis, ESG classification, FLS classification. This section outlines the tools and libraries, environment, model architecture and workflow used during experimentation.

#### **4.3.1 Testing Parameters**

For the initial evaluation of the model’s performance, the FinBERT model was used without further fine-tuning and was tested in its pretrained form as provided by Yang et al. (2020). The reason for this was to assess the baseline generalisability of the model on Malaysian financial texts, which may contain linguistic and contextual differences from the original US-based dataset used to initially train the models. The architecture of the models — a BERT-base encoder fine-tuned for financial domain classification — remained unchanged. Instead, the focus was on inference-only evaluation.

#### **4.3.2 Tools and Libraries Used**

The testing and evaluation process of the FinBERT model was conducted using Python in a cloud-based environment using Google Colab. The following tools and libraries were used during implementation of model testing and the evaluation pipeline:

**Table 14.**  
*Tools and Libraries*

| Library                   | Description                                                                                                                  | Code Snippet                                                                                                         |
|---------------------------|------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|
| Hugging Face Transformers | Provides general-purpose architecture for NLP and was used to load the pretrained FinBERT models and tokenise the sentences. | <pre>!pip install transformers from transformers import BertTokenizer, BertForSequenceClassification, pipeline</pre> |
| Pandas                    | A data analysis library used to load the Excel test set, control data structures and save results.                           | <pre>!pip install pandas import pandas as pd</pre>                                                                   |
| OpenPyXL                  | Intergrated with pandas to support reading and writing to Excel files (.xlsx)                                                | <pre>!pip install openpyxl</pre>                                                                                     |

### 4.3.3 Testing Workflow

The testing workflow for evaluating the three pretrained FinBERT-2020 models was designed to follow a structured and repeatable sequence of steps to ensure consistency in processing, ease the process of tracing results and to establish compatibility with later stages in fine-tuning. The overall testing process can be broken down into five key phases:

**Figure 22.**  
*Workflow Overview*

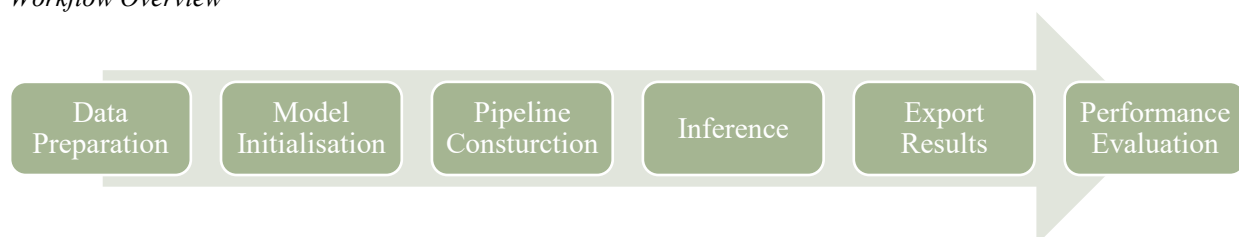


Figure 22 illustrates the workflow overview of the testing phase. The phase began with ‘Data Preparation’, whereby the test set, compiled in an Excel spreadsheet, was loaded using the pandas and openpyxl library. Each task (sentiment analysis, ESG classification, FLS classification) has its own corresponding sheet labelled accordingly (“SA”, “ESG”, “FLS”) and each sheet contains two key columns: one for one for the sentence IDs (“ID”) and one for the actual text data (“Sentences”). For consistency, the same set of 613 sentences were used for testing.

**Figure 23.**  
*Code Snippet for Task Preparation (Sentiment Analysis)*

```
df = pd.read_excel("Testset.xlsx", sheet_name="SA")
sentences = df['Sentences'].tolist()[:613]
ids = df['ID'].tolist()
```

Following this, in the ‘Model Initialisation’ step, the required pretrained FinBERT models were initialised using the Hugging Face transformers library. Three distinct versions of the FinBERT model were employed in this study, each trained for a different classification task relevant to financial text analysis:

- i. **yiyanghkust/finbert-tone** for sentiment analysis, which categorises sentences into positive, neutral, or negative tone.
- ii. **yiyanghkust/finbert-esg** for ESG classification, used to identify whether the content relates to Environmental, Social, Governance factors, or if there are none.
- iii. **yiyanghkust/finbert-fls** for forward-looking statements classification, which is used to classify sentences to predict the business outlook of an institution.

**Figure 24.**  
*Code Snippet for Sentiment Analysis*

```
finbert = BertForSequenceClassification.from_pretrained('yiyanghkust/finbert-tone', num_labels=3)
tokenizer = BertTokenizer.from_pretrained('yiyanghkust/finbert-tone')
```

**Figure 25.***Code Snippet for ESG Classification*

```
finbert = BertForSequenceClassification.from_pretrained('yiyanghkust/finbert-esg', num_labels=4)
tokenizer = BertTokenizer.from_pretrained('yiyanghkust/finbert-esg')
```

**Figure 26.***Code Snippet for FLS Classification*

```
finbert = BertForSequenceClassification.from_pretrained('yiyanghkust/finbert-fls', num_labels=3)
tokenizer = BertTokenizer.from_pretrained('yiyanghkust/finbert-fls')
```

Each model was loaded with the appropriate number of classification labels using the `BertForSequenceClassification` class and accompanied by its corresponding tokenizer via `BertTokenizer`.

Next, the ‘Pipeline Construction’ step was conducted to simplify the model’s inference. This was done by creating a pipeline object each task using the pipeline function from the transformers library, which bundles together the model and tokenizer into a high-level utility for text classification. The same pipeline interface is reused with all three models making the testing logic consistent across experiments.

**Figure 27.***Code Snippet for Pipeline Execution (Sentiment Analysis)*

```
from transformers import BertTokenizer, BertForSequenceClassification, pipeline

nlp = pipeline("text-classification", model=finbert, tokenizer=tokenizer)
```

By making use of the built-in tokenization and classification methods, large batches of sentences were able to be processed efficiently.

During the ‘Inference’ step, the test sentences were passed through the classification pipeline in batches and each sentence was processed and produced a ‘label’ to indicate the predicted class (i.e. “positive”, “Environmental”, “Specific-FLS”) and a ‘score’ which

indicated the model's confidence in its prediction. A batch size of 32 was utilised to improve the model's performance and to reduce memory usage during this step.

**Figure 28.**

*Code Snippet for Batch Classification of Test Set Sentences*

```
results = nlp(sentences, batch_size=32)
```

This process was repeated independently for all three models to generate three different prediction sets from the same input sentences.

The final step is 'Export Results' where the predicted results for each classification task were stored in a structured format using a pandas dataframe.

**Figure 29.**

*Code Snippet for Dataframe Conversion*

```
#dataframe conversion
results_df = pd.DataFrame(output)
results_df.to_excel('FinBERT_Test_Results_with_ID.xlsx', index=False)
```

The results were exported as Excel files. The resulting output included four columns: 'ID' referring to the original identifier assigned during the test set preparation phase; 'Sentence', containing the original input sentence; 'Predicted Label', which contained the model's prediction and 'Score', representing the model's confidence in its prediction.

To complete the testing phase, the model's performance was measured in the 'Performance Evaluation' step. The model's predictions were first imported from the Excel output file (FinBERT\_SA\_Test\_Results.xlsx) using the pandas library, and the first 613 rows were selected to maintain consistency with the prepared test set.

**Figure 30.**

*Code Snippet for Performance Metrics Library*

```
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
import seaborn as sns
import matplotlib.pyplot as plt
```

**Figure 31.***Code Snippet for Importing and Preparing Prediction Results for Evaluation*

```
df = pd.read_excel("FinBERT_SA_Test_Results.xlsx")
df = df.iloc[:613]
```

Within the Excel file, the predicted classification, its confidence score as well as the ground truth annotations were stored in separate columns which eased the process of drawing comparisons between the model's outputs and the ground truth.

**Figure 32.***Structure of Excel File*

|   | A       | B                                                                                                                                                                                                                                                                                           | C                   | D        | E                  |
|---|---------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|----------|--------------------|
| 1 | ID      | Sentence                                                                                                                                                                                                                                                                                    | Predicted Sentiment | Score    | Sentiment Analysis |
| 2 | MB188   | Under the initiative, Maybank Islamic created opportunities for social enterprises to engage with Maybank employees and the public and promote awareness about the role social enterprises have to play in sustainable social development.                                                  | Positive            | 0.983836 | Positive           |
| 3 | CIMB16  | Work together with other banks, clients, governments and policymakers and support each other in developing capabilities and the necessary methodologies to measure climate impact, align with global and local climate goals, and help transition to low-carbon, climate-resilient economy. | Neutral             | 0.948773 | Neutral            |
| 4 | BNM1241 | Our advocacy aimed to ensure the CMIM continues to reflect the needs of the region.                                                                                                                                                                                                         | Neutral             | 0.976133 | Neutral            |
| 5 | BNM611  | This facilitates effective monetary policy transmission as adjustments to loan repayments will take effect following an OPR change.                                                                                                                                                         | Positive            | 0.972808 | Neutral            |
| 6 | BNM239  | A prospective borrower could also be offered suggestions to improve his or her financial position, such as ways to generate more income or rationalise existing debt and other commitments before reapplying.                                                                               | Positive            | 0.985315 | Neutral            |
| 7 | BNM325  | While this is a positive development and a sign of progress in the country's health outcomes, this development naturally contributes to higher medical costs.                                                                                                                               | Positive            | 1        | Negative           |
| 8 | BNM1126 | Key Payment Trends<br>Usage of e-payment services continued to gain further traction.                                                                                                                                                                                                       | Positive            | 1        | Positive           |

To ensure data consistency, the actual and predicted labels under the 'Sentiment Analysis' and 'Predicted Sentiment' columns were pre-processed—this included converting them to lowercase and removing leading or trailing spaces. Any rows that did not match the valid sentiment labels ('positive', 'neutral', 'negative') were filtered out, resulting in a clean subset (df\_clean) for evaluation.

**Figure 33.***Label Cleaning and Filtering*

```
valid_sentiments = ['positive', 'neutral', 'negative']

df['Sentiment Analysis'] = df['Sentiment Analysis'].astype(str).str.strip().str.lower()
df['Predicted Sentiment'] = df['Predicted Sentiment'].astype(str).str.strip().str.lower()

df_clean = df[
    df['Sentiment Analysis'].isin(valid_sentiments) &
    df['Predicted Sentiment'].isin(valid_sentiments)
]
```

**Figure 34.**

*Code Snippet for Performance Metrics Evaluation and Confusion Matrix Plotting*

```
y_true = df_clean['Sentiment Analysis']
y_pred = df_clean['Predicted Sentiment']

#accuracy
accuracy = accuracy_score(y_true, y_pred)
print("Accuracy:", accuracy)

#confusion matrix
cm = confusion_matrix(y_true, y_pred, labels=valid_sentiments)
print("Confusion Matrix:\n", cm)

#plotting the confusion matrix
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
            xticklabels=valid_sentiments, yticklabels=valid_sentiments)
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")
plt.show()

#generating classification report
print("\nClassification Report:\n", classification_report(y_true, y_pred, labels=valid_sentiments))
```

The ground truth labels (`y_true`) and the model's predicted labels (`y_pred`) were then extracted. The `accuracy_score` function from the `scikit-learn` library was used to compute the overall accuracy of the model, reflecting the proportion of correct predictions out of the total number of test instances.

A confusion matrix was generated using `confusion_matrix`, enabling a detailed look into the distribution of correct and incorrect classifications across each sentiment category. This matrix was visualised with a heatmap using the `seaborn` library, where each cell represented the count of predictions compared to the true labels. Axis labels and a title were added to enhance clarity.

Finally, a classification report was generated using `classification_report`, which included precision, recall, and F1-score for each sentiment class. These metrics allowed for a more granular understanding of the model's strengths and weaknesses, particularly how well it identified positive, neutral, and negative tones in financial text.

## 4.4 Fine-Tuning Workflow

Once the performance of the existing FinBERT models was evaluated, further experimentation was conducted by fine-tuning the models on a labelled dataset. The objective of this process was to improve the model’s performance in task-specific environments – particularly for sentiment analysis, ESG classification, and FLS classification – by adapting the model to local linguistic nuances and contextual differences found in Malaysian financial reports.

### 4.4.1 Training Data Selection

The training dataset consists of 2,147 sentences which represent 70% of the MTFC dataset. The sentences were accompanied by the gold-standard annotations provided by the industry expert. These annotations served as the ground truth labels for the three classification tasks: sentiment analysis (positive, neutral, negative), ESG classification (environmental, social, governance), and forward-looking statement (FLS) classification (specific-FLS, non-specific-FLS, not-FLS).

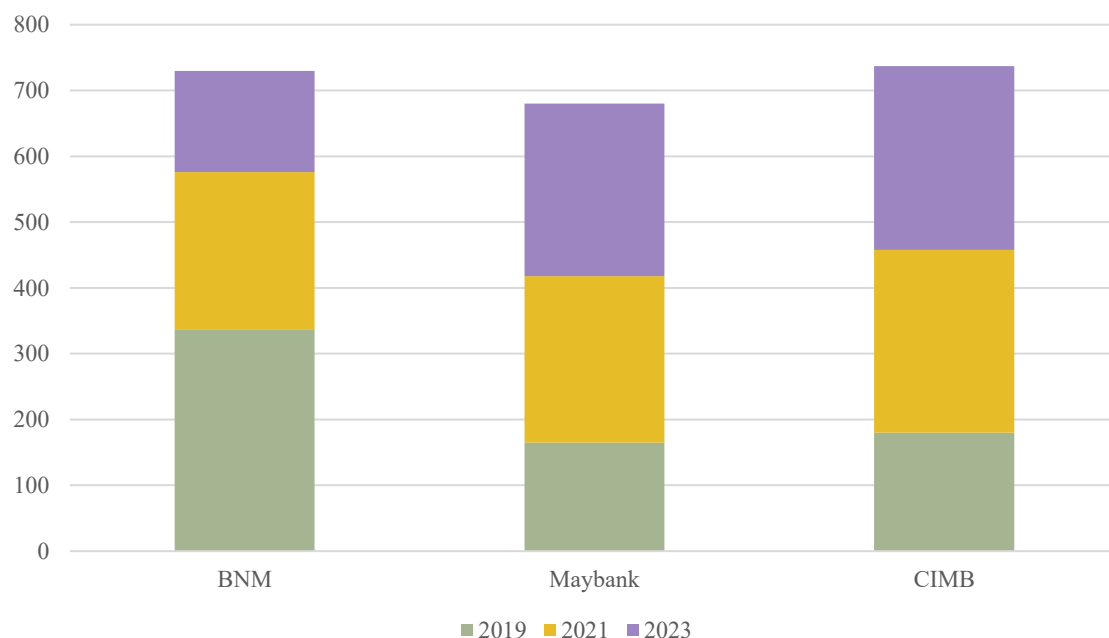
**Table 15.**  
*Dataset Composition for Model Fine-Tuning*

| Institution                         | 2019 | 2021 | 2023 | Total |
|-------------------------------------|------|------|------|-------|
| Bank Negara Malaysia<br>(BNM)       | 336  | 240  | 154  | 730   |
| Malayan Banking Berhad<br>(Maybank) | 165  | 253  | 262  | 680   |

|                                   |            |            |            |              |
|-----------------------------------|------------|------------|------------|--------------|
| Commerce International            |            |            |            |              |
| Merchant Bankers Berhad<br>(CIMB) | 180        | 278        | 279        | <b>737</b>   |
| <b>Total</b>                      | <b>681</b> | <b>771</b> | <b>695</b> | <b>2,147</b> |

Table 15 displays a breakdown of the training dataset used to fine-tune the FinBERT-2020 models across different institutions and years. A total of 2,147 sentences from three key institutions were used in the fine-tuning process. Of the 2,147 sentences, BNM contributed 336 sentences in 2019, 240 in 2021, and 154 in 2023, amounting to a total of 730 sentences; Maybank contributed 165 sentences in 2019, 253 in 2021, and 262 in 2023, totalling 680 sentences; and CIMB contributed 180 sentences in 2019, 278 in 2021, and 279 in 2023, for a total of 737 sentences.

**Figure 35.**  
*Breakdown of Sentence Contributions by Institution and Year*



Based on the figure above, it was observed that CIMB contributed the most number of sentences with 737 sentences while the year 2021 had the highest number of sentences amongst

the three years at 771 sentences. In comparison, 2019 had 681 sentences and 2023 had 695 sentences. Despite these variations, the dataset remains relatively balanced across each institution and year, and each bank is adequately represented in the dataset. This allows the FinBERT model to generalise more effectively across different time periods and financial contexts.

#### 4.4.2 Parameters

A set of carefully selected hyperparameters were used during the fine-tuning process of the FinBERT model. These parameters were applied consistently across all three classification tasks to allow fair comparison of the model outcomes. The configuration was implemented using the Hugging Face Trainer API as detailed in Figure 36.

**Figure 36.**

*Code Snippet for Fine-Tuning Parameters (FinBERT FLS)*

```
training_args = TrainingArguments(  
    output_dir='finbert-checkpoints',  
    run_name='finbert-fls',  
    eval_strategy='epoch',  
    save_strategy='epoch',  
    per_device_train_batch_size=32,  
    per_device_eval_batch_size=32,  
    num_train_epochs=5,  
    learning_rate=2e-5,  
    weight_decay=0.01,  
    load_best_model_at_end=True,  
    metric_for_best_model='accuracy',  
    logging_dir='./logs',  
    logging_strategy='epoch',  
    report_to='none'  
)
```

The figure above shows the training arguments implemented during the fine-tuning procedure. The batch size was set to 32 to balance memory efficiency and convergence speed whereas the learning rate ( $2e - 5$ ) used is a commonly used default for BERT-based models. A small weight decay of 0.01 was applied to help regularise the model and prevent overfitting.

The model was evaluated and saved at the end of each epoch, with the best performing model (based on accuracy) being saved for final evaluation.

Early stopping was also added with a patience value of 2, which terminates training if no improvement in validation accuracy is observed over two consecutive epochs. This step was implemented to reduce the possibility of overfitting while saving training time.

The performance metric used to monitor and guide checkpoint saving was accuracy, as it provides a clear measure for multi-class classification tasks. Logging was also performed at each epoch to monitor training and validation progress.

#### **4.4.3 Model Adaption and Training**

To begin the fine-tuning process, the pre-trained FinBERT model ([yiyanghkust/finbert-pretrain](#)) was used as the base model. It was further adapted to accommodate the specific label sets and class imbalance present in the MTFC dataset across the three classification tasks measured in this study: sentiment analysis, ESG classification, FLS classification. The process was implemented using the Hugging Face transformers library and carried out with the Trainer API with the following parameters:

The fine-tuning procedure begins with importing the necessary libraries, including Hugging Face's transformers, datasets and standard Python libraries such as pandas, numpy and sklearn as shown in Figure 37.

**Figure 37.***Code Snippet for Libraries*

```

import transformers
import torch
import pandas as pd
import numpy as np
from datasets import Dataset
from transformers import EarlyStoppingCallback
from sklearn.utils.class_weight import compute_class_weight
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from transformers import AutoConfig, BertTokenizer, BertForSequenceClassification, TrainingArguments, Trainer

```

Next, to address class imbalances in the training dataset, class weights were calculated using the `compute_class_weight` function from the `sklearn` library. This method calculates the inverse frequency of each class, giving heavier weights (importance) to underrepresented labels. The resulting weights were later incorporated into the model's loss function to reduce the possibility of biases towards the majority classes as well as to improve overall classification performance across all categories.

The model, training and validation datasets are then loaded from Excel files, and the columns are renamed for consistency.

**Figure 38.***Code Snippet for Model, Class Weights and Dataset Loading*

```

model_name = 'yiyanghkust/finbert-pretrain'

df = pd.read_excel('Training_Set(FLS).xlsx')
df_fls = df.rename(columns={'Sentences': 'sentence', 'FLS': 'label'})[['sentence', 'label']]
le = LabelEncoder()
df_fls['label'] = le.fit_transform(df_fls['label'])

labels = df_fls['label'].to_numpy()
class_weights_np = compute_class_weight(class_weight='balanced', classes=np.unique(labels), y=labels)
class_weights = torch.tensor(class_weights_np, dtype=torch.float)

df_val = pd.read_excel('Validation_Set(FLS).xlsx')
df_fls_val = df_val.rename(columns={'Sentences': 'sentence', 'FLS': 'label'})[['sentence', 'label']]
df_fls_val['label'] = le.transform(df_fls_val['label'])

```

Figure 38 displays the data preparation process, including the loading and formatting of the training and validation datasets, label encoding, and the computation of class weights to

address label imbalance. The training dataset was imported from an Excel file (Training\_Set(FLS).xlsx) and reformatted to include only the relevant columns: the input sentences and their corresponding FLS labels. To ensure compatibility with model training, the label values were encoded into numeric form using LabelEncoder. Similarly, the validation dataset (Validation\_Set(FLS).xlsx) was loaded and processed using the same label encoding scheme to ensure consistency during evaluation.

To address the class imbalances within the training dataset, the `compute_class_weight` function is implemented with a ‘balance’ strategy. These weights were then converted into a PyTorch tensor and later applied to the model’s loss function to reduce bias toward majority classes.

For the class weights to be applied during training, a custom subclass `BertForSequenceClassification` was implemented.

**Figure 39.**  
*Code Snippet for Custom BERT Subclass Implementation*

```
class WeightedBertForSequenceClassification(BertForSequenceClassification):
    def __init__(self, config, class_weights):
        super().__init__(config)
        self.class_weights = class_weights

    def forward(self, input_ids=None, attention_mask=None, token_type_ids=None, labels=None):
        outputs = super().forward(input_ids, attention_mask, token_type_ids)
        logits = outputs.logits

        loss = None
        if labels is not None:
            loss_fct = torch.nn.CrossEntropyLoss(weight=self.class_weights.to(logits.device))
            loss = loss_fct(logits, labels)

        return (loss, logits) if loss is not None else logits

config = AutoConfig.from_pretrained(model_name, num_labels=len(le.classes_))
model = WeightedBertForSequenceClassification.from_pretrained(model_name, config=config, class_weights=class_weights)
```

The figure above shows the process of implementing the custom BERT subclass. The subclass is designed to override the forward method to introduce a weighted cross-entropy loss function, which ensures that misclassifications of minority classes are penalised more heavily

during training, allowing for better model sensitivity to underrepresented labels. This approach was consistently applied across all three classification tasks.

Next, the pre-trained tokenizer (BertTokenizer) from the FinBERT model is used to tokenize the text data.

**Figure 40.**

*Code Snippet for BERT Tokenizer*

```
tokenizer = BertTokenizer.from_pretrained(model_name)

def tokenize_function(examples):
    tokens = tokenizer(examples['sentence'], truncation=True, padding='max_length', max_length=128)
    tokens['label'] = examples['label']
    return tokens

dataset_train = Dataset.from_pandas(df_fls)
dataset_train = dataset_train.map(tokenize_function, batched=True)
dataset_train.set_format(type='torch', columns=['input_ids', 'token_type_ids', 'attention_mask', 'label'])

dataset_val = Dataset.from_pandas(df_fl_val)
dataset_val = dataset_val.map(tokenize_function, batched=True)
dataset_val.set_format(type='torch', columns=['input_ids', 'token_type_ids', 'attention_mask', 'label'])
```

Based on the figure above, each sentence in the training and validation sets was tokenized and this also involved truncation and a padding set to a maximum length of 128 tokens to maintain consistency in input dimensions. The tokenized datasets were then converted into the Hugging Face Dataset format and prepared for training by setting the appropriate tensor columns (input\_ids, token\_type\_ids, attention\_mask, and label).

Once the tokenization and dataset formatting were completed, the training parameters outlined in Section 4.4.2 were applied.

**Figure 41.**

*Code Snippet for Trainer Initialisation with Early Stopping*

```
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=dataset_train,
    eval_dataset=dataset_val,
    compute_metrics=compute_metrics,
    callbacks=[EarlyStoppingCallback(early_stopping_patience=2)]
)
```

The training process was executed using the Trainer class from Hugging Face's Transformers library and was initialized with the fine-tuned model, training arguments, tokenized training and validation datasets, and a custom evaluation function based on accuracy. To mitigate overfitting, an early stopping mechanism was also implemented, halting training if no improvement in validation performance was observed over two consecutive epochs.

Finally, the `train()` method is called to start the fine-tuning process as shown in Figure 42.

**Figure 42.**  
*Executing the Training Loop*

```
trainer.train()
```

During the training process, the model was evaluated on the validation set after each epoch, and accuracy was used to track its performance. Once training was completed, the newly fine-tuned model's performance was assessed using the test set previously outlined in Section 4.2.3.

**Figure 43.**  
*Code Snippet for Model Evaluation (Post-Fine Tuning)*

```
df = pd.read_excel("Testset.xlsx", sheet_name="FLS")
sentences = df['Sentences'].tolist()[:613]
labels = df['FLS Classification'].tolist()[:613]
ids = df['ID'].tolist()[:613]

labels_encoded = le.transform(labels)

df_test = pd.DataFrame({
    'sentence': sentences,
    'label': labels_encoded,
    'ID': ids
})

dataset_test = Dataset.from_pandas(df_test)

def tokenize_function(examples):
    return tokenizer(examples['sentence'], truncation=True, padding='max_length', max_length=128)

dataset_test = dataset_test.map(tokenize_function, batched=True)
dataset_test.set_format(type='torch', columns=['input_ids', 'token_type_ids', 'attention_mask', 'label'])

model.eval()
predictions = trainer.predict(dataset_test)
print(predictions.metrics)
```

Figure 43 details the method used to evaluate the fine-tuned model. This test set (Testset.xlsx) remained independent of the training and validation set used during fine-tuning and consists of 613 annotated sentences. The test data was pre-processed in the same way as the training data, including label encoding, tokenization, and dataset formatting. The model was set to evaluation mode (model.eval()), and predictions were obtained using the Trainer.predict() method. The resulting evaluation metrics, such as accuracy, were printed and used to assess the performance improvements gained through fine-tuning.

To further evaluate the performance of the fine-tuned models, a more detailed analysis was conducted using classification metrics and a confusion matrix.

**Figure 44.**

*Code Snippet for Classification Report and Confusion Matrix Plotting (Post Fine-Tuning)*

```
y_true = dataset_test['label'].numpy()
predictions = trainer.predict(dataset_test)
y_pred = np.argmax(predictions.predictions, axis=1)
valid_labels = le.classes_

accuracy = accuracy_score(y_true, y_pred)
print("Accuracy:", accuracy)

cm = confusion_matrix(y_true, y_pred, labels=range(len(valid_labels)))
print("Confusion Matrix:\n", cm)
sns.heatmap(cm, annot=True, fmt='d', cmap='Greens',
            xticklabels=valid_labels, yticklabels=valid_labels)
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")
plt.show()

print("\nClassification Report:\n", classification_report(y_true, y_pred, target_names=valid_labels))
```

The true labels (y\_true) are extracted from the test dataset and the model's predictions are then obtained using the Trainer.predict() method. The predicted class labels (y\_pred) are determined by selecting the index of the highest predicted probability for each instance. The overall accuracy of model's performance was also calculated using accuracy\_score and a confusion matrix was generated using a Seaborn heatmap (sns.heatmap) to visualize the distribution of true versus predicted labels. Finally, a classification report consisting of the

accuracy, precision, recall, and F1-score was printed using `classification_report` for each class label defined in the label encoder.

#### 4.5 Model Deployment via Gradio Interface

In order to conduct interactive testing and real-time demonstration of the fine-tuned FinBERT models, a Gradio-based interface was developed and deployed. Gradio is a Python library used primarily to create simple, interactive and functional web applications to host machine learning models.

The process began by importing the necessary libraries to the Google Colab environment.

**Figure 45.**  
*Code Snippet for Library Installation and Environment Setup*

```
!pip install transformers gradio openpyxl --quiet

import torch
from transformers import AutoTokenizer, AutoModelForSequenceClassification
import pandas as pd
import gradio as gr
import matplotlib.pyplot as plt
import io
```

Figure 45 shows the required libraries used for model deployment, which includes the transformers library and its accompanying components designed to simplify working with pre-trained transformer models like BERT. Several supporting packages are also installed – pandas for data manipulation and analysis; gradio for interface deployment; and matplotlib for generating data visualisations. `AutoModelForSequenceClassification` was used to load the three fine-tuned models as well as a shared tokenizer (`AutoTokenizer`). For accessibility and compatibility purposes, the model was executed on the CPU as shown in Figure 46.

**Figure 46.**  
*CPU Configuration for Model Inference*

```
#force CPU usage
device = torch.device("cpu")
```

Following this, the fine-tuned models are loaded from a personal HuggingFace repository (hannell/).

**Figure 47.**  
*Loading Fine-Tuned FinBERT Models and Tokenizer from Hugging Face*

```
#load models from Hugging Face
sentiment_model = AutoModelForSequenceClassification.from_pretrained("hannell/finbertMY-sentiment").to(device)
esg_model = AutoModelForSequenceClassification.from_pretrained("hannell/finbertMY-esg").to(device)
fls_model = AutoModelForSequenceClassification.from_pretrained("hannell/finbertMY-fls").to(device)
tokenizer = AutoTokenizer.from_pretrained("hannell/finbertMY-sentiment")
```

Next, label mapping dictionaries, which associate each model's output with an accompanying label that is readable to humans, are defined.

**Figure 48.**  
*Label Mapping and Classification Function Definition*

```
#label mapping
label_maps = {
    "sentiment": {0: "Negative", 1: "Neutral", 2: "Positive"},
    "esg": {0: "Environmental", 1: "Social", 2: "Governance", 3: "None"},
    "fls": {0: "Specific FLS", 1: "Non-Specific FLS", 2: "Not FLS"},
}

def classify(text, model, label_map):
    inputs = tokenizer(text, return_tensors="pt", truncation=True, padding=True).to(device)
    with torch.no_grad():
        outputs = model(**inputs)
    probs = torch.nn.functional.softmax(outputs.logits, dim=-1)[0]
    label = label_map[torch.argmax(probs).item()]
    return label, probs.tolist()
```

Figure 48 shows the various mappings implemented during this process. Separate mappings are determined for each classification task and are used during inference to translate the predicted class indices into output labels on the Gradio interface.

**Figure 49.***Code Snippet for Single Sentence Processing*

```
#single sentence processing
def process_input(text, model, label_map):
    if not text.strip():
        return "Please enter a valid sentence.", pd.DataFrame(columns=["Sentence", "Prediction", "Confidence"])

    label, probs = classify(text, model, label_map)
    score = round(max(probs), 4)

    row = pd.DataFrame([{"Sentence": text,
                        "Prediction": label,
                        "Confidence": score}])

    return f"Prediction: {label}", row
```

Figure 49 details the classification process for single sentence inputs. Before the sentence is passed through the `classify()` function, the input text is first checked to determine if it is in the correct form i.e., not empty or just whitespace. If the sentence is valid, then it passes through the function and a predicted label and corresponding confidence score is obtained. A DataFrame stores the sentence, predicted label, confidence value as well as the highest probability score, which was rounded. The function then returns a structured prediction summary and table to be displayed on the interface.

**Figure 50.***Code Snippet for Excel File Processing*

```
#excel file processing
def process_excel(file_obj, model, label_map):
    try:
        df = pd.read_excel(file_obj)

        if not any(col.lower() == "sentences" for col in df.columns):
            return pd.DataFrame([{"Error": "Excel file must contain a column named 'Sentences'"}]), None

        col_name = next(col for col in df.columns if col.lower() == "sentences")
        sentences = df[col_name].dropna().tolist()

        rows = []
        for sent in sentences:
            inputs = tokenizer(sent, return_tensors="pt", truncation=True, padding=True)
            outputs = model(**inputs)
            probs = torch.nn.functional.softmax(outputs.logits, dim=-1)[0]
            pred = torch.argmax(probs).item()
            label = label_map[pred]
            score = round(probs[pred].item(), 4)
            rows.append({"Sentence": sent, "Prediction": label, "Confidence": score})

        result_df = pd.DataFrame(rows)
        return result_df, None
    except Exception as e:
        return pd.DataFrame([{"Error": f"Error reading Excel file: {str(e)}"}]), None
```

Similar to Figure 49, Figure 50 details the same core process with a few additional steps. This section is designed to handle batch input through an Excel file upload. The pandas library first reads the uploaded file and then determines if a column named ‘Sentences’ exists within the file. If the column is present, it proceeds to extract text data, tokenise each sentence and pass the sentences through the selected model for classification. Like the single sentence processing, a DataFrame stores the predicted label and the confidence score for every entry. However, if the ‘Sentences’ column is not found within the Excel file, an error message is returned.

Finally, all components are integrated into an interactive Gradio interface as shown in Figure 51:

**Figure 51.**  
*Code Snippet for Gradio Interface Build*

```
with gr.Blocks() as demo:
    gr.Markdown("## FinBERTMY")

    with gr.Tab("Sentiment Analysis"):
        gr.Markdown("**Detect financial sentiment: Positive, Neutral, or Negative**")

        #single sentence input
        with gr.Row():
            sentiment_text = gr.Textbox(label="Enter a sentence")
            sentiment_btn = gr.Button("Analyze Sentence")
            sentiment_result = gr.Textbox(label="Prediction Summary")
            sentiment_table = gr.Dataframe(
                label="Classification Results",
                headers=["Sentence", "Prediction", "Confidence"],
                interactive=False
            )

        #file upload
        with gr.Row():
            sentiment_file = gr.File(label="Upload Excel File",
file_types=[".xlsx"])
            sentiment_file_btn = gr.Button("Analyze File")
            sentiment_file_table = gr.Dataframe(
                label="Classification Results",
                headers=["Sentence", "Prediction", "Confidence"],
                interactive=False,
            )

        #sentence button
        sentiment_btn.click(
```

```

        fn=lambda txt: process_input(txt, sentiment_model,
label_maps["sentiment"]),
        inputs=sentiment_text,
        outputs=[sentiment_result, sentiment_table]
    )

    #file button
    sentiment_file_btn.click(
        fn=lambda f: process_excel(f, sentiment_model,
label_maps["sentiment"])[0],
        inputs=sentiment_file,
        outputs=sentiment_file_table
    )

    with gr.Tab("ESG Classification"):
        gr.Markdown("**Classify ESG: Environmental, Social, Governance**")

        #single sentence input
        with gr.Row():
            esg_text = gr.Textbox(label="Enter a sentence")
            esg_btn = gr.Button("Analyze Sentence")
            esg_result = gr.Textbox(label="Prediction Summary")
            esg_table = gr.Dataframe(
                label="Classification Results",
                headers=["Sentence", "Prediction", "Confidence"],
                interactive=False
            )

        #file upload
        with gr.Row():
            esg_file = gr.File(label="Upload Excel File",
file_types=[".xlsx"])
            esg_file_btn = gr.Button("Analyze File")
            esg_file_table = gr.Dataframe(
                label="Classification Results",
                headers=["Sentence", "Prediction", "Confidence"],
                interactive=False,
            )

        #sentence button
        esg_btn.click(
            fn=lambda txt: process_input(txt, esg_model, label_maps["esg"]),
            inputs=esg_text,
            outputs=[esg_result, esg_table]
        )

        #file button
        esg_file_btn.click(
            fn=lambda f: process_excel(f, esg_model, label_maps["esg"])[0],
            inputs=esg_file,
            outputs=esg_file_table
        )

    with gr.Tab("FLS Classification"):
        gr.Markdown("**Classify Forward-Looking Statements: Specific-FLS, Non-specific FLS, Not-FLS**")

        #single sentence input
        with gr.Row():

```

```

        fls_text = gr.Textbox(label="Enter a sentence")
        fls_btn = gr.Button("Analyze Sentence")
        fls_result = gr.Textbox(label="Prediction Summary")
        fls_table = gr.Dataframe(
            label="Classification Results",
            headers=["Sentence", "Prediction", "Confidence"],
            interactive=False
        )

    #file upload
    with gr.Row():
        fls_file = gr.File(label="Upload Excel File",
file_types=[".xlsx"])
        fls_file_btn = gr.Button("Analyze File")
        fls_file_table = gr.Dataframe(
            label="Classification Results",
            headers=["Sentence", "Prediction", "Confidence"],
            interactive=False,
        )

    #sentence button
    fls_btn.click(
        fn=lambda txt: process_input(txt, fls_model, label_maps["fls"]),
        inputs=fls_text,
        outputs=[fls_result, fls_table]
    )

    #file button
    fls_file_btn.click(
        fn=lambda f: process_excel(f, fls_model, label_maps["fls"])[0],
        inputs=fls_file,
        outputs=fls_file_table
    )

demo.launch(share=True)

```

## 4.6 Chapter Summary

This chapter outlined the implementation phase, which included dataset preparation, the experiment setup for baseline performance evaluation and model fine-tuning as well as the performance evaluation criteria. An analysis of the model's performance pre fine-tuning and post fine-tuning will be conducted in the subsequent chapter.

## Chapter 5: Performance Evaluation and Results Analysis

### 5.1 Introduction

This chapter summarises and concludes the experimentation phase of this study, which included evaluating and fine-tuning the FinBERT models for financial text analysis. The evaluation focuses on the effectiveness of the models (pretrained and fine-tuned) in performing three classification tasks: sentiment analysis, ESG classification, and FLS classification. The performance of each model is studied based on standard classification metrics, and the training procedure is also reviewed and assessed. A sentence-level qualitative analysis is also conducted to gain deeper insights into the model's predictions and their contextual relevance. This chapter concludes with discussing the limitations, practical applications and recommendations for future improvements.

### 5.2 Model Training Overview

The fine-tuning process was carried out on a Google Colab environment using the HuggingFace Trainer API. Prior to training, the NumPy library was reinstalled to ensure compatibility with the FinBERT model configurations and other dependent packages using `!pip install numpy==1.26.4 --force-reinstall`. This step was done to ensure any version-related conflicts during model loading and tokenization were prevented.

All three FinBERT models were trained using the same hyperparameters for compatibility purposes across the classification tasks measured. The training process implemented several arguments: a learning rate of  $2e - 5$ , a batch size of 32 and weight decay of 0.01 over a maximum of 5 epochs.

The training outcomes and output results of each of the models are illustrated below in Figures 52, 53 and 54 and Tables 16, 17, and 18 respectively:

**Figure 52.***Training and Validation Loss with Accuracy Across Epochs for Sentiment Analysis Model*


| Epoch | Training Loss | Validation Loss | Accuracy |
|-------|---------------|-----------------|----------|
| 1     | 0.805100      | 0.662688        | 0.622150 |
| 2     | 0.536200      | 0.639260        | 0.690554 |
| 3     | 0.381300      | 0.661830        | 0.719870 |
| 4     | 0.273200      | 0.711616        | 0.710098 |
| 5     | 0.214700      | 0.720940        | 0.729642 |

**Table 16.***Training Output for Sentiment Analysis Model*

| Metric                     | Value               |
|----------------------------|---------------------|
| Global Step                | 340                 |
| Training Loss              | 0.44208665735581343 |
| Training Runtime (seconds) | 15068.5527          |
| Samples Per Second         | 0.712               |
| Steps Per Second           | 0.023               |
| Total FLOPs                | 706,130,634,827,520 |
| Epochs Completed           | 5                   |

The sentiment analysis model concluded training after 5 epochs and recorded a training loss of 0.442. The entire process took 15,068 seconds, which is approximately 4 hours and 11 minutes long, with a total of 340 training steps completed. The average number of samples was 0.712 per second while the step rate was 0.023 steps per second, reflecting the performance constraints of CPU-based processing. Lastly, the total estimated floating-point operations (FLOPs) was recorded at 706 trillion.

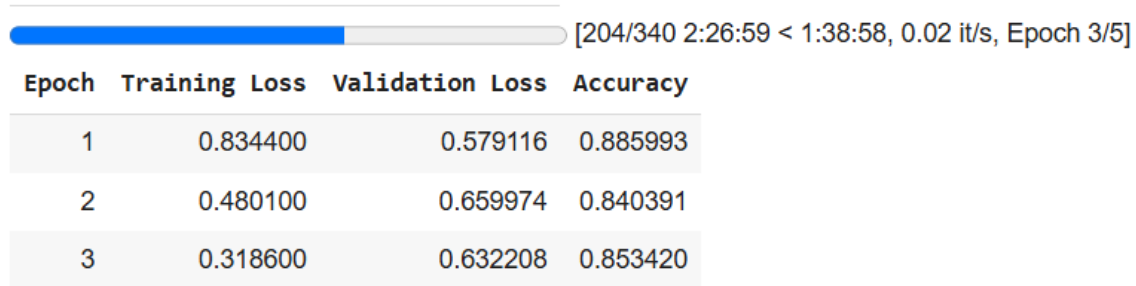
**Figure 53.***Training and Validation Loss with Accuracy Across Epochs for ESG Classification Model*


| Epoch | Training Loss | Validation Loss | Accuracy |
|-------|---------------|-----------------|----------|
| 1     | 1.061400      | 0.759752        | 0.726384 |
| 2     | 0.633000      | 0.626338        | 0.739414 |
| 3     | 0.432500      | 0.579635        | 0.755700 |
| 4     | 0.301100      | 0.590965        | 0.771987 |
| 5     | 0.232400      | 0.600468        | 0.778502 |

**Table 17.***Training Output for ESG Classification Model*

| Metric                     | Value               |
|----------------------------|---------------------|
| Global Step                | 340                 |
| Training Loss              | 0.5320894746219411  |
| Training Runtime (seconds) | 14865.839           |
| Samples Per Second         | 0.722               |
| Steps Per Second           | 0.023               |
| Total FLOPs                | 706,136,974,832,640 |
| Epochs Completed           | 5                   |

For the ESG classification model, training was concluded 5 epochs, achieving a training loss of 0.532. The training time was 12,865 seconds long (4 hours and 8 minutes) with a total of 340 training steps completed. The sample throughput was slightly higher at 0.722 samples per second, with a step rate of 0.023 steps per second. The total FLOPs were approximately 706 trillion, which makes the sentiment analysis model as a result of similar model complexity and dataset structure.

**Figure 54.***Training and Validation Loss with Accuracy Across Epochs for FLS Classification Model***Table 18.***Training Output for FLS Classification Model*

| Metric                     | Value               |
|----------------------------|---------------------|
| Global Step                | 204                 |
| Training Loss              | 0.5443581693312701  |
| Training Runtime (seconds) | 8868.2596           |
| Samples Per Second         | 1.21                |
| Steps Per Second           | 0.038               |
| Total FLOPs                | 423,678,380,896,512 |
| Epochs Completed           | 3                   |

The FLS classification completed training in 3 epochs, even though it uses the same dataset as the previous two model for training, halting early due to the early stopping mechanism in place. The final training loss was 0.544 and the training time was significantly shorter at 8,868 seconds (approximately 2 hours and 28 minutes). A total of 204 training steps were completed, processing at 1.21 samples per seconds and 0.038 steps per second. The total FLOPs recorded was 423 trillion.

Model checkpoints were saved based on the best-performing epoch, determined by the evaluation loss and macro F1-score on the validation set across all tasks, which was implemented using the `load_best_model_at_end=True` and `metric_for_best_model='accuracy'` arguments. The entire training process was closely monitored to prevent overfitting and maximise generalisability on unseen data. At the

end, each model was successfully optimised for its respective classification task, thereby forming the basis for the following evaluation.

### 5.3 Performance Evaluation (Pre Fine-Tuning)

A baseline performance evaluation of each pretrained FinBERT model was conducted before proceeding with model fine-tuning. The evaluation focused on three key tasks—sentiment analysis, ESG classification, and forward-looking statement (FLS) classification—each of which plays a significant role in understanding corporate communication and narrative strategy. 613 sentences from the curated test set were evaluated across several performance metrics, namely accuracy, precision, recall, and F1-score and a confusion matrix was plotted to visualise the distribution of correct and incorrect predictions across different classes.

#### 5.3.1 Sentiment Analysis Results

The FinBERT model achieved an overall accuracy of 61.17%, indicating moderate effectiveness in correctly classifying the sentiment of financial texts.

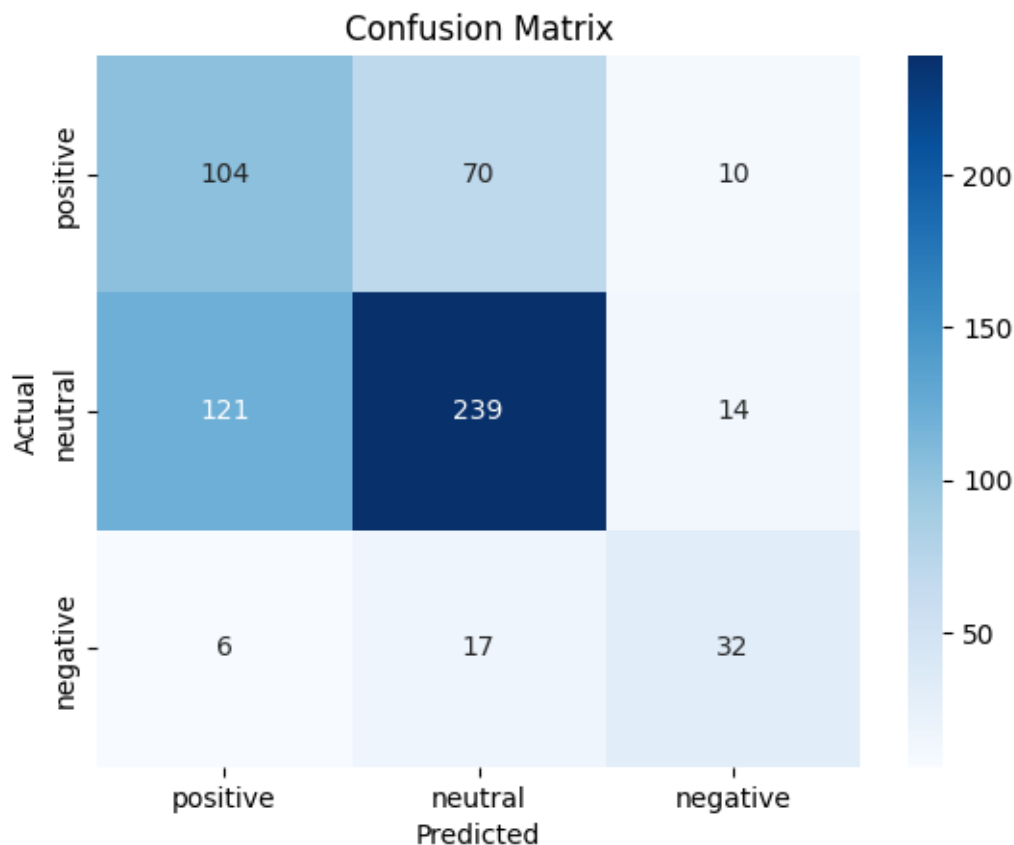
**Table 19.**  
*Classification Report for Sentiment Analysis*

| Class            | Precision | Recall | F1-Score | Support |
|------------------|-----------|--------|----------|---------|
| Positive         | 0.45      | 0.57   | 0.50     | 184     |
| Neutral          | 0.73      | 0.64   | 0.68     | 374     |
| Negative         | 0.57      | 0.58   | 0.58     | 55      |
| Accuracy         |           |        | 0.61     | 613     |
| Macro Average    | 0.58      | 0.60   | 0.59     | 613     |
| Weighted Average | 0.63      | 0.61   | 0.62     | 613     |

Based on the results obtained, the neutral class received the highest F1-score (0.68), which reflects the model’s advanced ability in identifying sentences with a neutral tone, a common trait in financial reporting. Conversely, positive sentiments scored the lowest during

the precision evaluation at 0.45, indicating that a significant portion of the non-positive sentences were misclassified as positive statements. This also suggests that the model struggled to distinguish positive sentiments from other classes, particularly neutral, which may reflect overlapping linguistic patterns or insufficient representation of positive samples during training.

**Figure 55.**  
*Confusion Matrix for Sentiment Analysis Classification*



Based on the confusion matrix produced in Figure 55, it can be concluded that the model tended to confuse positive and neutral sentences as 70 sentences were misclassified as neutral. A substantial number of neutral sentences (121) were also misclassified as positive, further highlighting the challenge in differentiating between these two classes.

However, despite these limitations, the model performed fairly well when detecting negative sentences and managed to maintain a balance between the precision and recall metrics (both measured at 0.58). This indicates that while the model shows potential, there is room for

improvement, particularly in fine-tuning the detection of positive tone in nuanced financial language.

### 5.3.2 ESG Classification Results

The performance of the FinBERT model on the ESG classification task was evaluated across four labels: environmental, social, governance, and none. Overall, the model achieved an accuracy of 32.79%, which is relatively low. This suggests that the model struggles to correctly classify ESG-related content into the correct categories.

**Table 20.**  
*Classification Report for ESG Classification.*

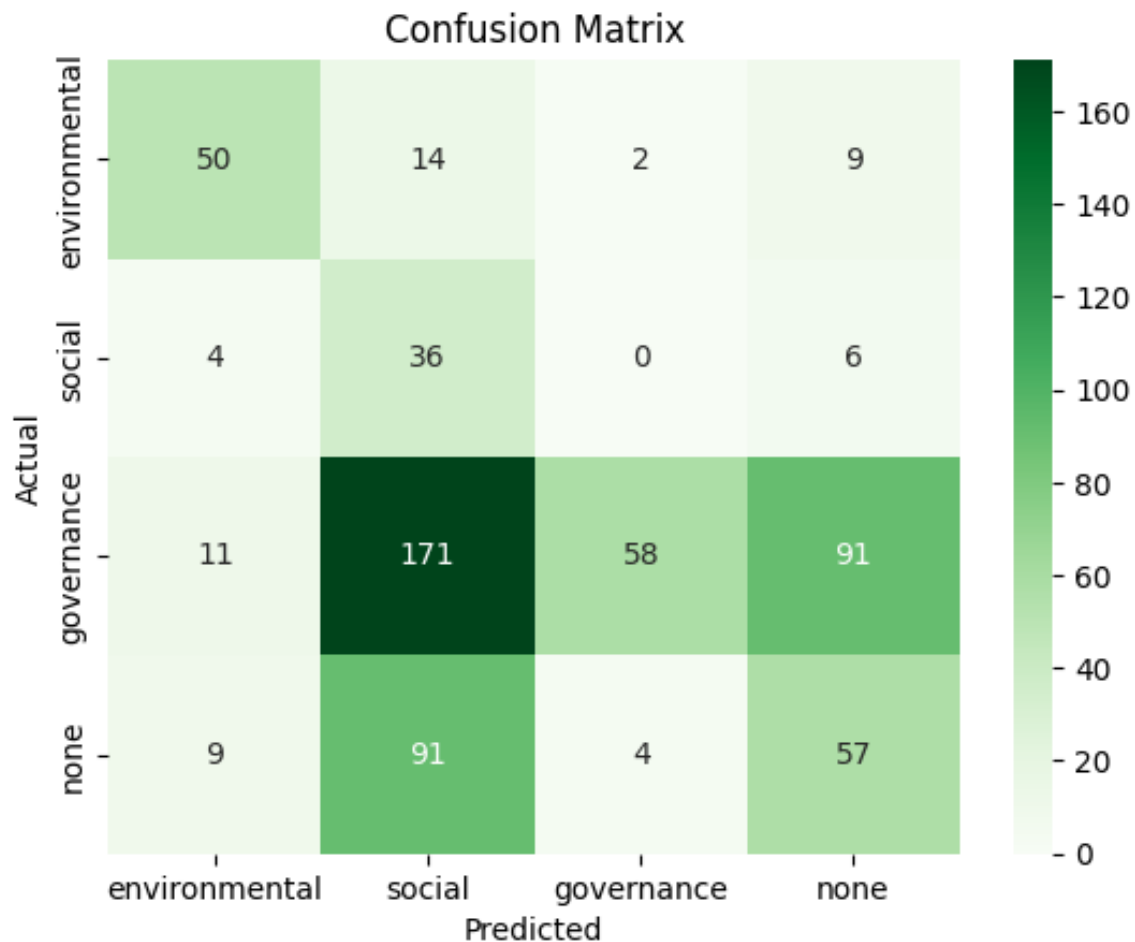
| Class            | Precision | Recall | F1-Score | Support |
|------------------|-----------|--------|----------|---------|
| Environmental    | 0.68      | 0.57   | 0.67     | 75      |
| Social           | 0.12      | 0.78   | 0.20     | 46      |
| Governance       | 0.91      | 0.18   | 0.29     | 331     |
| None             | 0.35      | 0.35   | 0.35     | 161     |
| Accuracy         |           |        | 0.33     | 613     |
| Macro Average    | 0.51      | 0.49   | 0.38     | 613     |
| Weighted Average | 0.67      | 0.33   | 0.35     | 613     |

The macro-average F1-score and weighted-average F1-score measured at 0.38 and 0.35 respectively, which further highlights weak performance across classes.

The model’s poor performance, particularly regarding the governance class, suggests that the model is over-predicting the social class. This can be observed from the high number of governance and none instances that were misclassified as social. Additionally, an imbalance in the dataset — with governance being the dominant class — may also have contributed to this. Furthermore, the presence of the ‘none’ label, which is inherently ambiguous, may have

caused overlaps with other ESG themes, thereby confusing the model and adding to the misclassifications.

**Figure 56.**  
*Confusion Matrix for ESG Classification*



From the confusion matrix plotted, it can be observed that a total of 171 governance sentences were misclassified as social or none, which contributed to the low recall result (0.18) for governance, indicating that the model frequently fails to detect this class. In contrast, the social class had high recall results (0.78) with 36 out of 46 sentences being correctly classed.

**Table 21.***Sample of Misclassifications by FinBERT model for ESG Classification.*

| ID     | Sentence                                                                                                                                                                                                                    | Predicted Label | True Label    |
|--------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|---------------|
| CIMB73 | The subsidiaries that have achieved more than 30% female representation on the Board are CIMB Bank Berhad (40%), and CIMB Islamic Bank Berhad (33%).                                                                        | None            | Governance    |
| BNM196 | With this framework, Islamic financial institutions are encouraged to nurture their customers, investors and other stakeholders to adopt more sustainable practices.                                                        | Social          | Governance    |
| MB76   | The Bank encourages constructive engagements with customers on the merits of sustainable practices.                                                                                                                         | Governance      | Environmental |
| MB67   | The ESG Policy takes into account global standards set by the International Finance Corporation as well as UN Human Rights Policy.                                                                                          | Social          | Governance    |
| BNM10  | As the effect of monetary policy takes at least two or three quarters to be felt in the economy, the MPC, in making its decisions, assesses and deliberates on the outlook for both domestic economic growth and inflation. | None            | Governance    |

Table 21 displays a few samples of misclassifications that occurred during testing. It is clear that these arise not from clear model errors, but from sentence ambiguity and the multidisciplinary nature of ESG content. For example, MB67 discusses the ESG Policy alignment with global standards set by the International Finance Corporation and UN Human Rights Policy. While the reference to policy suggests a strong governance theme, the model

predicted it as social, which is likely due to the mention of *human rights* – a common theme that often falls in the social domain of ESG. This example highlights the overlap between governance frameworks and social impact, making it challenging for a single-label classification model to assign the most appropriate category.

### 5.3.3 Forward-Looking Statements (FLS) Classification Results

The FinBERT model achieved an overall accuracy of 86.62%, indicating relatively high performance in correctly classifying FLS sentences. However, a closer look at the classification report and confusion matrix reveals important nuances across the individual classes.

**Table 22.**  
*Classification Report for FLS Classification.*

| Class            | Precision | Recall | F1-Score | Support |
|------------------|-----------|--------|----------|---------|
| Specific FLS     | 0.36      | 0.79   | 0.49     | 24      |
| Non-specific FLS | 0.21      | 0.11   | 0.14     | 47      |
| Not FLS          | 0.95      | 0.94   | 0.94     | 542     |
| Accuracy         |           |        | 0.87     | 613     |
| Macro Average    | 0.50      | 0.61   | 0.52     | 613     |
| Weighted Average | 0.87      | 0.87   | 0.86     | 613     |

The macro-average showed moderate performance with precision measuring at 0.50, recall at 0.61 and F1 at 0.52. This metric usually treats all the classes equally; therefore, these results reflect the challenge faced by the model to classify underrepresented classes. On the other hand, the weighted average recorded 0.87 for precision, 0.87 for recall and 0.86 for F1. The high metric results may be due to the heavy influence of the Not FLS class but this also causes class-performance issues to be masked.

**Figure 57.**  
*Confusion Matrix for FLS Classification*

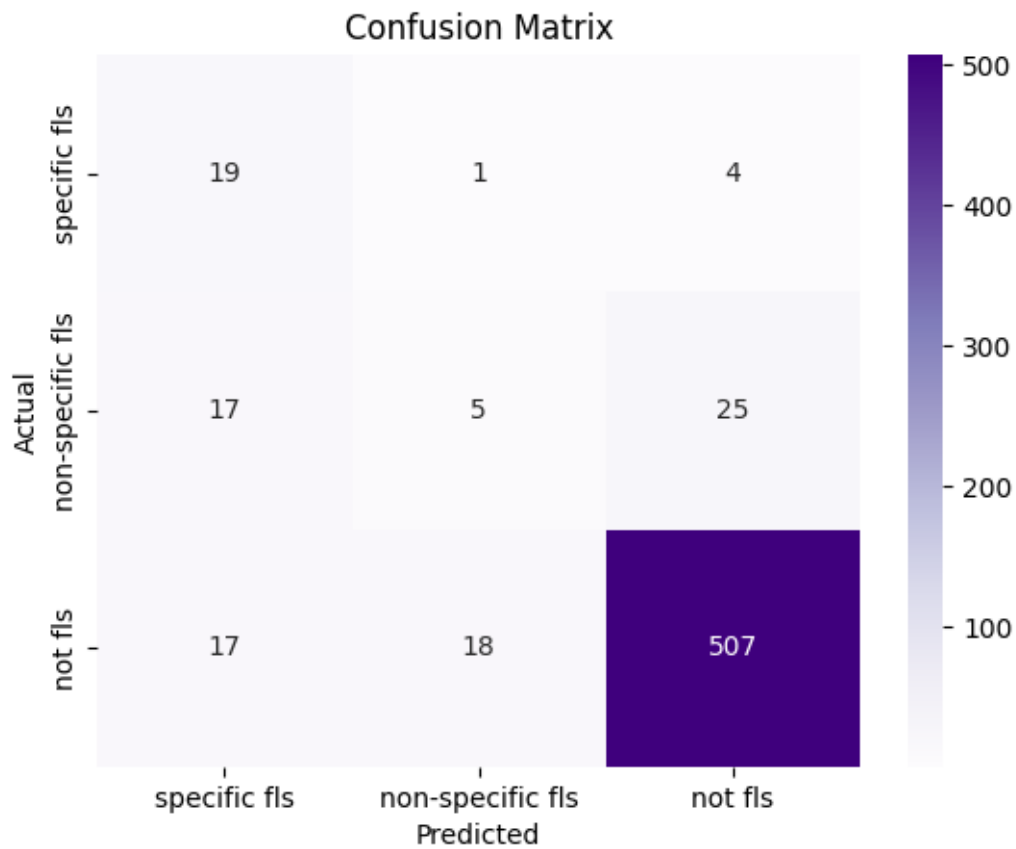


Figure 57 shows the confusion matrix plotted to evaluate the model. Not FLS sentences recorded the highest precision and recall at 95% and 94% respectively and dominated the overall test set with 507 sentences being classified. This heavily influenced the overall accuracy score and also suggests that the model was biased toward predicting the majority class.

Non-Specific FLS sentences showed the weakest performance with only 5 out of 47 sentences being correctly classified. Its recall measured at 11% while the precision result was also relatively low at 21%. This suggests that the model struggled to distinguish vague or general forward-looking statements, often confusing them with either specific FLS or non-FLS.

## 5.4 Performance Evaluation (Post Fine-Tuning)

Once fine-tuning was completed, the performance of the newly fine-tuned FinBERT models was evaluated across all three classification tasks. The same performance metrics used to evaluate the baseline performance were utilised here to ensure uniformity. Comparisons were drawn to assess the improvements made after fine-tuning, reflecting the model's ability to adapt to contextual and linguistic aspects of financial text.

### 5.4.1 Sentiment Analysis Results

The fine-tuned FinBERT model for sentiment analysis achieved an overall accuracy of 71.78%, meaning that it was able to correctly classify 72% of all instances across the 613 test samples. The accuracy of the model in correctly classifying sentiments increased by 10.60%. This suggests that fine-tuning was able to enhance the model's ability to capture domain-specific sentiment nuances in financial texts.

**Table 23.**  
*Classification Report for Sentiment Analysis Post Fine-Tuning.*

| Class            | Precision | Recall | F1-Score | Support |
|------------------|-----------|--------|----------|---------|
| Positive         | 0.58      | 0.62   | 0.60     | 184     |
| Neutral          | 0.82      | 0.75   | 0.78     | 374     |
| Negative         | 0.61      | 0.80   | 0.69     | 55      |
| Accuracy         |           |        | 0.72     | 613     |
| Macro Average    | 0.67      | 0.72   | 0.69     | 613     |
| Weighted Average | 0.73      | 0.72   | 0.72     | 613     |

Table 23 details the classification report of the sentiment analysis task after fine-tuning. Based on the report, it was observed that Positive sentiment class had a precision of 0.58 and a recall of 0.62, resulting in an F1-score of 0.60. These results indicate that while the model was moderately efficient at identifying positive sentences. On the other hand, for the Negative sentiment, precision was 0.61 and recall was significantly higher at 0.80, which produced an

F1-score of 0.78, suggesting that the model was good at capturing most negative instances. However, the slightly lower precision also points to some non-negative sentences being misclassified as negative. The Neutral class showed the strongest performance amongst the three classes, with precision measuring at 0.82, recall at 0.75 and F1-score at 0.78. This shows that the model was most confident and consistent in classifying neutral sentiments, though this may be due to the class's dominance or the clarity of neutral tones within the dataset.

The macro average's precision measured at 0.67, recall at 0.72 and F1-score at 0.69. This reflects the model's balanced performance across all classes regardless of their frequency. The weighted average, on the other hand, which accounts for the number of samples in each class, recorded a slightly higher precision (0.73) and F1-score (0.72). This can be attributed to the larger influence of the Neutral class in the overall metrics.

**Figure 58.**  
*Confusion Matrix for Sentiment Analysis Post Fine-Tuning*

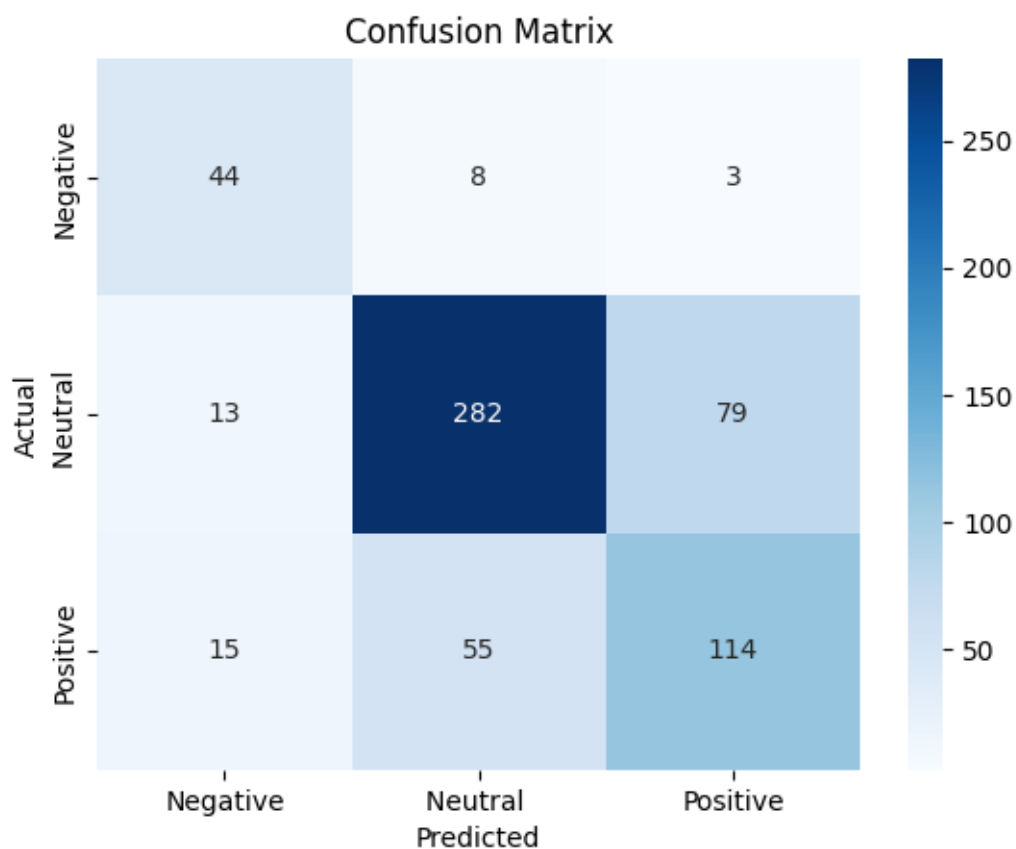


Figure 58 illustrates the confusion matrix plotted based on the performance of the model after fine-tuning. Out of all the classes, Neutral was the best predicted class with 282 correct predictions out of 374 test samples. However, 79 of these sentences were misclassified as Positive. The Negative class had the smallest test sample (55) but performed relatively well with 44 correct predictions, recording only 11 misclassifications. Conversely, the Positive class had a bigger test sample with 184 sentences but had a larger number of sentences that were misclassified as Neutral (55), showing that the model tends to default to Neutral. However, the model did correctly classify 144 sentences as Positive.

#### 5.4.2 ESG Classification Results

After being fine-tuned on a curated dataset, FinBERT model for ESG classification achieved an overall accuracy of 71.12%, an increase of about 38.42%. The model showed notable improvements across all performance metrics.

**Table 24.**  
*Classification Report for ESG Classification Post Fine-Tuning*

| Class            | Precision | Recall | F1-Score | Support |
|------------------|-----------|--------|----------|---------|
| Environmental    | 0.75      | 0.84   | 0.79     | 75      |
| Social           | 0.59      | 0.65   | 0.62     | 46      |
| Governance       | 0.83      | 0.73   | 0.77     | 331     |
| None             | 0.35      | 0.35   | 0.35     | 161     |
| Accuracy         |           |        | 0.71     | 613     |
| Macro Average    | 0.68      | 0.71   | 0.69     | 613     |
| Weighted Average | 0.73      | 0.71   | 0.72     | 613     |

From the results shown in Table 24, it was observed that the best performing class was Governance which recorded a precision of 0.83, recall of 0.73. producing an F1-score of 0.77, indicating that when the Governance class is predicted by the model, it is highly likely to be correct. However, the slightly lower recall indicates that some Governance-related sentences

may have been missed. Conversely, the worst performing class was None, with all of its metrics measuring at 0.35. This implies that the model struggled to differentiate sentences that did not belong to any ESG label and likely classified them into one of the three classes. Both false negatives and false positives for this class are high.

The model performed strongly when classifying sentences in the Environmental class, achieving a precision of 0.75 and a high recall of 0.84. The F1-score was then recorded at 0.79, which indicates a favourable balance between precision and recall, suggesting that the model was highly efficient at detecting and correctly labelling sentences that fell into the Environmental domain with minimal misclassifications. In contrast, while the model performed moderately for the Social class – with precision and recall valuing at 0.59 and 0.65 respectively – the F1-score measured at 0.62, indicating that although the model can detect Social content to some extent, it struggles with precision and occasionally mislabels non-Social content as Social.

Next, the macro average had a precision of 0.68, recall of 0.71 and F1-score of 0.72, demonstrating that the model maintains a relatively consistent performance across all classes. In contrast, the weighted average results confirm that while the model performs reliably on more frequent classes, such as Governance and None, it struggles to classify underrepresented or more ambiguous ones as it achieved 0.73 in precision, 0.71 in recall and 0.72 for F1-score.

**Figure 59.**  
*Confusion Matrix for ESG Classification Post Fine-Tuning*

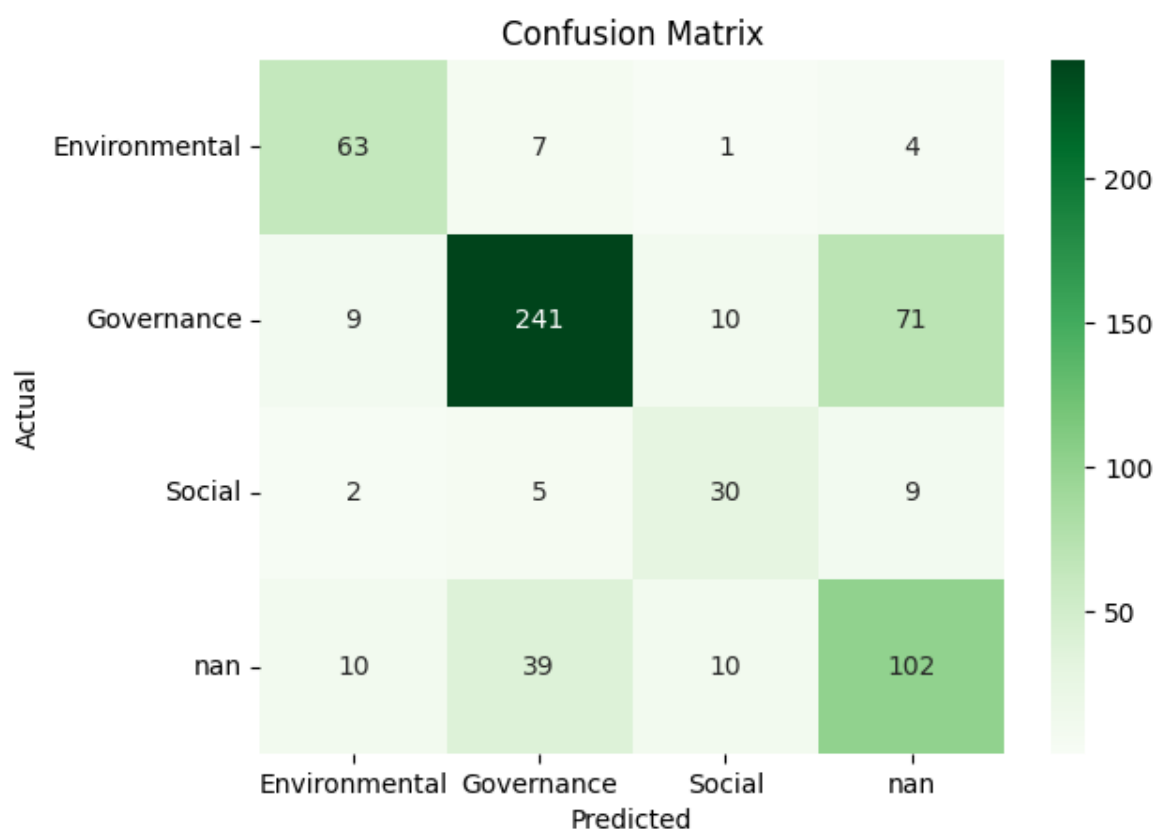


Figure 59 illustrates the confusion matrix plotting the performance of the FinBERT model after fine-tuning. It was observed that out of all the classes, Environment had the best performance with 63 correctly predicted sentences out of 75 samples. There were, however, 12 misclassified sentence (mostly Governance) but overall, the model performs relatively well. The class with the lowest performance was None, with only 102 out of 161 sentences being correctly predicted. 10 sentences were misclassified as Environmental, 39 as Governance and 10 as Social. This shows that the model struggled to clearly distinguish None from ESG-related categories, in particular sentences that fell into the Governance domain. These results also indicate that non-ESG content may share similar language or context with Governance related sentences, which may have led to the frequent confusion.

### 5.4.3 Forward-Looking Statements (FLS) Classification Results

An overall accuracy of 88.74% was achieved post fine-tuning by the model for FLS classification task, indicating that the model is highly effective in distinguishing between forward-looking and non-forward-looking content. Generally, it was observed that the model performed strongly across all classes but several challenges in classification can be observed.

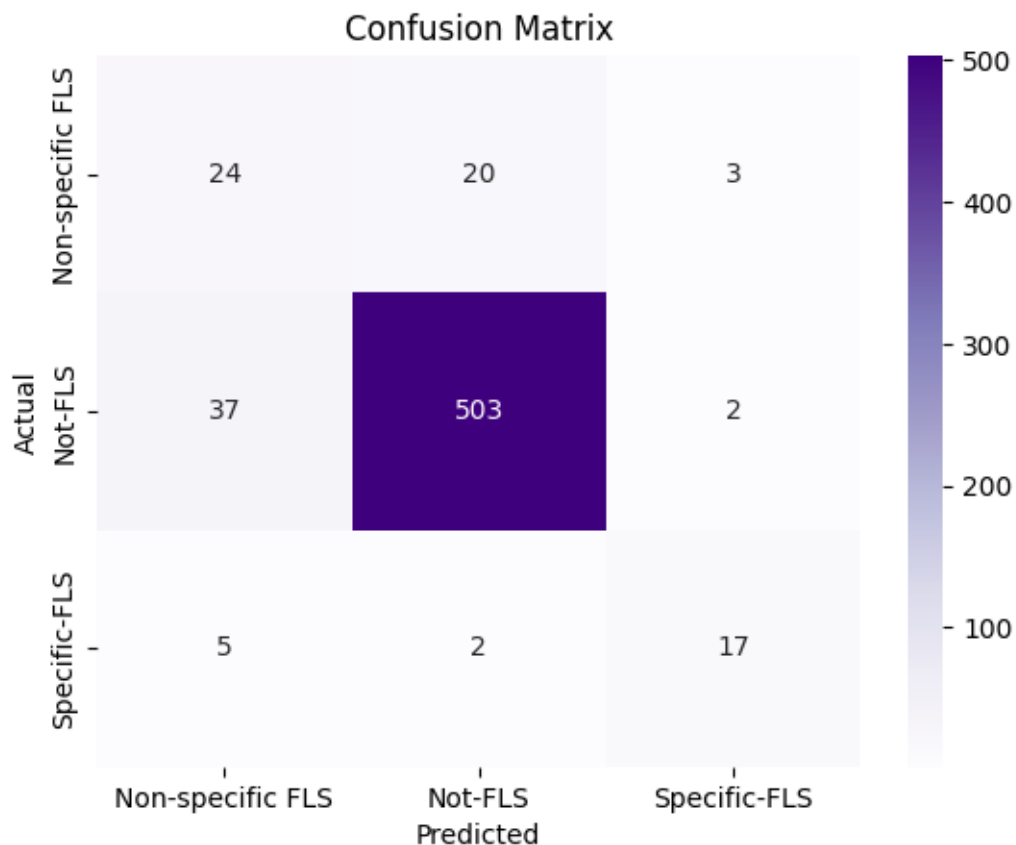
**Table 25.**  
*Classification Report for FLS Classification Post Fine-Tuning.*

| Class            | Precision | Recall | F1-Score | Support |
|------------------|-----------|--------|----------|---------|
| Specific FLS     | 0.77      | 0.71   | 0.74     | 24      |
| Non-specific FLS | 0.36      | 0.51   | 0.42     | 47      |
| Not FLS          | 0.96      | 0.93   | 0.94     | 542     |
| Accuracy         |           |        | 0.89     | 613     |
| Macro Average    | 0.70      | 0.72   | 0.70     | 613     |
| Weighted Average | 0.91      | 0.89   | 0.90     | 613     |

The table above depicts the results recorded by the model following fine-tuning for the FLS classification task. The strongest performance was observed for the Not-FLS class, with a precision of 0.96, recall of 0.93 and, F1-score of 0.94, which can be attributed to the high number of samples present in the dataset (support = 542). In contrast, despite having significantly lower representation within the dataset (support = 24), the Specific-FLS class achieved a precision of 0.77, recall of 0.71 and an F1-score of 0.74, suggesting a strong ability to identify clearly stated forward-looking statements. On the other hand, the Non-specific FLS class performance was relatively weak, with precision being recorded at 0.36, recall at 0.51, and F1-score at 0.42. These results suggest that the model struggled to recognise vague or implied FLS sentences.

The macro average measured a precision of 0.70, recall of 0.72 and F1-score of 0.70, reflecting the model’s overall capability across classes. Additionally, the weighted average achieved solid results with 0.91 for precision, 0.89 in recall and an F1-score of 0.90, indicating a strong bias towards the majority class – which in this case is the Not FLS class.

**Figure 60.**  
*Confusion Matrix for FLS Classification Post Fine-Tuning*



From the figure above, the best performing class was Not-FLS, with 503 correctly classified sentences out of 542 samples. However, 37 sentences were misclassified as Non-specific FLS and another 2 as Specific-FLS, which suggests that while the model demonstrates strong overall accuracy for this class, there still exists some confusion with forward-looking content – particularly non-specific sentences that may share overlapping language patterns. In contrast, both the Non-specific FLS and Specific-FLS classes showed weaker performances. For the Non-specific FLS class, only 24 out of 47 sentences were correctly classified, with 20

being misclassified as Not-FLS and 3 as Specific-FLS. This suggests that the model has difficulty recognizing general forward-looking language and often confuses it with unrelated content. Similarly, out of 24 sentences in the Specific-FLS class, 17 were correctly predicted while the rest were mostly misclassified as Non-specific FLS, reflecting the challenge faced by the model to recognise general forward-looking language without confusing it with unrelated content.

## 5.5 Analysis of FinBERT Model Performance

The following section presents an analysis of the FinBERT model's performance before and after fine-tuning across the three classification tasks. The results of key evaluation metrics such as accuracy, precision, recall, and F1-score were studied in order to assess the extent of performance changes and identify areas where fine-tuning contributed the most to improved classification outcomes.

### 5.5.1 Comparative Analysis of Pre- and Post-Fine-Tuning Results

To better understand the significance of fine-tuning, the performance of the FinBERT model using selected evaluation metrics were recorded. The comparison highlights the differences between pre-training and fine-tuning using a curated dataset.

**Table 26.**  
*Summary of Metric Improvements Across Tasks Post Fine-Tuning*

| Task               | Metric                    | Pre Fine-Tuning | Post Fine-Tuning | Gain   |
|--------------------|---------------------------|-----------------|------------------|--------|
| Sentiment Analysis | Accuracy                  | 61.17%          | 78.92%           | 17.75% |
|                    | Macro F1-score            | 0.58            | 0.77             | 0.19   |
| ESG Classification | Macro Precision           | 0.35            | 0.58             | 0.23   |
|                    | Macro Recall              | 0.42            | 0.63             | 0.21   |
|                    | Macro F1-score            | 0.38            | 0.61             | 0.23   |
| FLS Classification | Recall (Non-specific FLS) | 0.11            | 0.51             | 0.40   |

|  |                             |      |      |      |
|--|-----------------------------|------|------|------|
|  | F1-score (Non-specific FLS) | 0.14 | 0.42 | 0.28 |
|--|-----------------------------|------|------|------|

Table 26 summarises the observed gains across several key metrics and provides an overview of how fine-tuning has influenced the model’s ability to accurately classify sentences within their respective domains across the three classification tasks.

For sentiment analysis, the most notable improvement was in accuracy (Gain = 17.75%) and macro F1-score (Gain = 0.19), indicating that the model’s performance across positive, neutral and negative sentiments was more balanced post fine-tuning. In the ESG classification task, substantial gains were observed across all macro-averaged metrics (precision, recall, F1-score), heavily suggesting improved detection of ESG related content. As for the FLS classification task, fine-tuning contributed to a marked increase in recall and F1-score for the Not-specific FLS class, which was previously underrepresented and posed a challenge for the model to accurately predict the class.

These results collectively demonstrate that fine-tuning not only improved the overall accuracy of all the models across their respective tasks but also enhanced the model’s capability to detect subtle and complex patterns in financial text. The improvements across multiple tasks reinforce the value of domain-specific fine-tuning for financial NLP applications.

The improvements of performance metrics highlighted are visualised in the graphs below:

**Figure 61.**  
*Sentiment Analysis: Pre vs Post Fine-Tuning Performance*

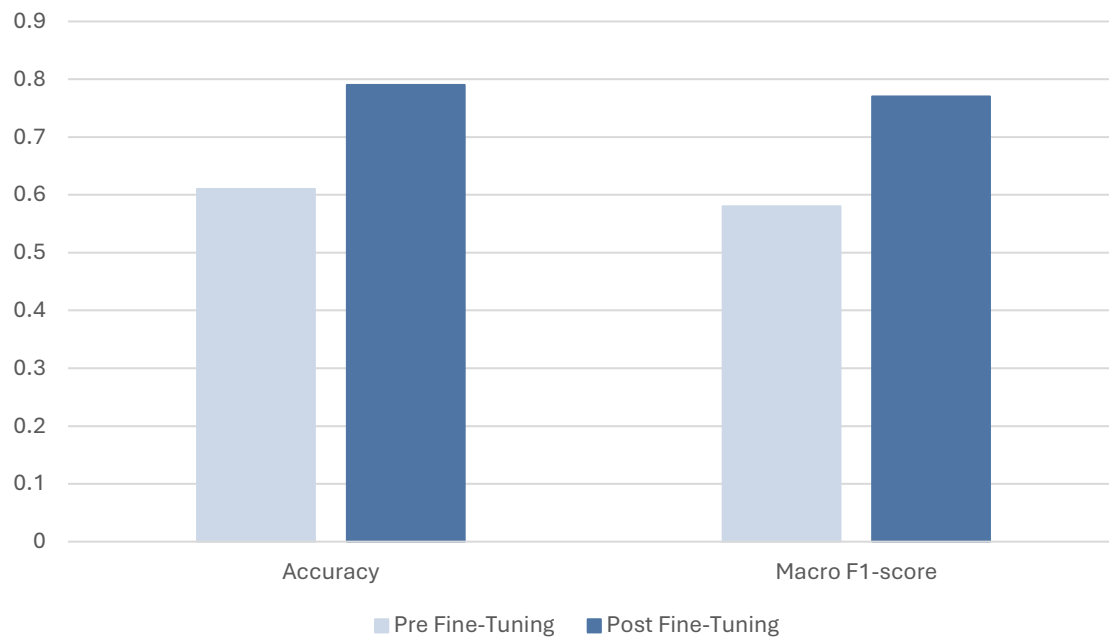


**Figure 62.**  
*ESG Classification: Pre vs Post Fine-Tuning Performance*

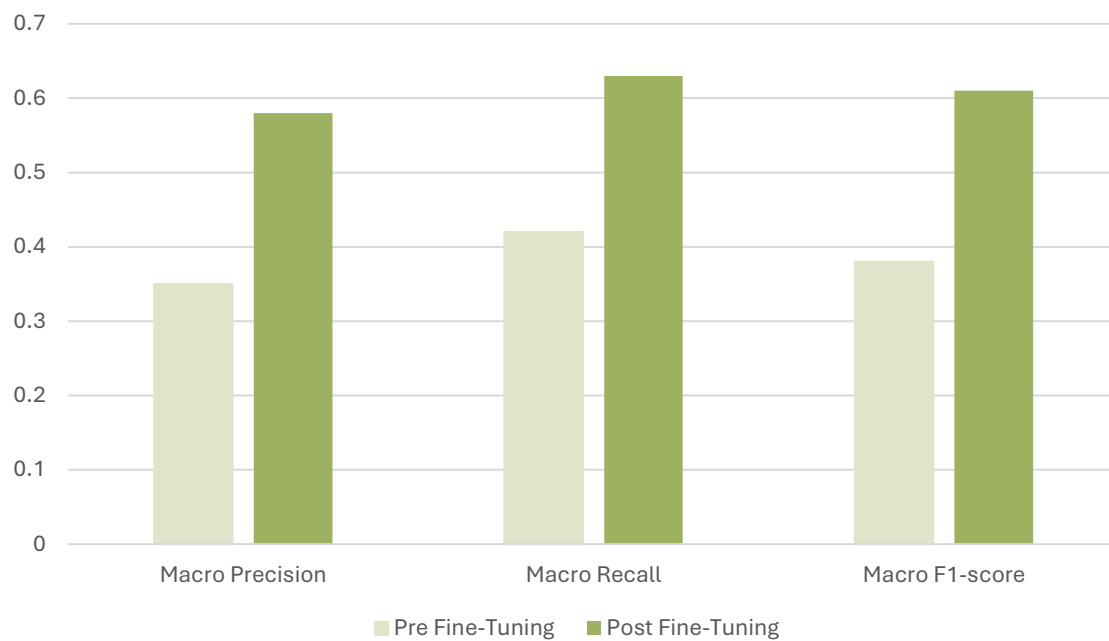
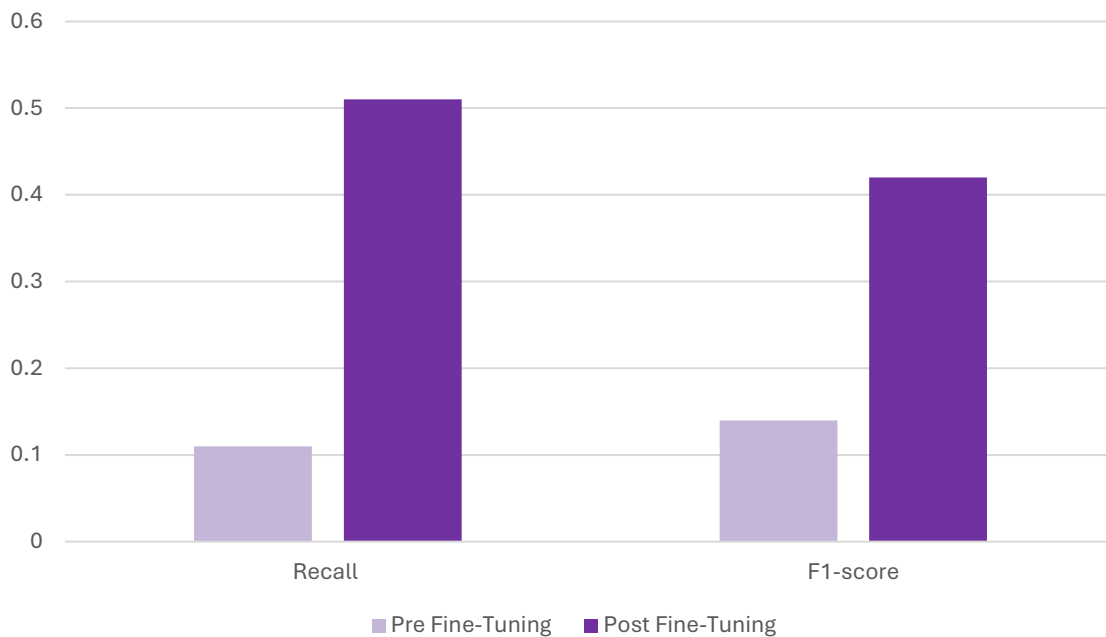


**Figure 63.**

*FLS Classification (Non-Specific FLS Class): Pre vs Post Fine-Tuning Performance*



### 5.5.2 Model Strengths and Weaknesses

The fine-tuned FinBERT models showcased distinct strengths and weaknesses throughout the evaluation process across all three classification tasks. As a domain-specific language model adapted to financial texts, FinBERT's performance greatly improved as a result of fine-tuning on annotated data and enhanced its ability to interpret contextually rich and semantically complex sentences. The evaluation revealed not only how well model captured linguistic patterns commonly found financial reporting but also uncovered where it struggled – particularly with ambiguous language, multi-class classifications and underrepresented categories. Together, these findings provide an in-depth understanding to the model's behaviour as well as its practical utility and areas where it may benefit from further refinement.

**Table 27.**  
*Summary of Model Strengths and Weaknesses by Classification Task (Post Fine-Tuning)*

| Task               | Strength                                                                                                                                        | Weakness                                                                                                                                  |
|--------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| Sentiment Analysis | <ul style="list-style-type: none"> <li>• Able to clearly identify sentiment cues in financial texts.</li> </ul>                                 | <ul style="list-style-type: none"> <li>• Occasional confusion between neutral and negative tones</li> </ul>                               |
|                    | <ul style="list-style-type: none"> <li>• High post fine-tuning accuracy and macro F1-score.</li> </ul>                                          | <ul style="list-style-type: none"> <li>• Struggled with implicit or factual sentiment</li> </ul>                                          |
| ESG Classification |                                                                                                                                                 | <ul style="list-style-type: none"> <li>• Experience significant difficulties with multi-label predictions and subtle ESG cues.</li> </ul> |
|                    | <ul style="list-style-type: none"> <li>• Good at recognising social/governance keywords.</li> <li>• Improved macro-averaged metrics.</li> </ul> | <ul style="list-style-type: none"> <li>• Lower precision for Governance label.</li> </ul>                                                 |
| FLS Classification | <ul style="list-style-type: none"> <li>• Strong performance on factual/historical content.</li> </ul>                                           | <ul style="list-style-type: none"> <li>• Struggled with vague or broadly worded future statements without timelines.</li> </ul>           |

Table 27 summarises the strengths and weaknesses of the FinBERT model post-tuning across each classification task, highlighting areas where the model performed well and where limitations were observed despite performance improvements.

In sentiment analysis, the model demonstrated a strong ability to identify sentiment polarity in financial narratives as seen in its high post fine-tuning accuracy (78.92%) and macro F1-score (0.77). The model excelled at detecting positive and neutral sentiments, which are common in financial texts. However, it struggled at times to distinguish between negative and neutral statements, leading to occasional misclassification.

For the ESG classification task, the model’s overall performance jumped post fine-tuning across most performance metrics. In particular the macro-average scores. However, while these metrics improved, lower precision and recall in minority classes indicate the impact of class imbalance and semantic complexity in multi-label classification.

Similarly, for the FLS classification task, the model was especially efficient in identifying the Not-FLS class. This suggests a reliable understanding of factual and historical, which is essential in filtering non-predictive statements in financial reporting. In spite of that, several weaknesses were also observed, particularly the struggle to recognise ambiguous forward-looking statements with no temporal or strategic markers present, which falls into the domain of the Non-specific FLS class.

These observations highlight the fine-tuned model’s ability to capture meaningful patterns in financial language, particularly for well-represented and clearly defined classes. Nonetheless, its limitations in handling ambiguous phrasing, multi-label complexities, and underrepresented categories point to areas for future improvement.

### **5.5.3 Qualitative Prediction Examples**

To further illustrate the performance of the fine-tuned FinBERT model, a qualitative analysis on sentence-level predictions was conducted. These examples offer insights into the model’s decision-making process, particularly in cases that may align well with human judgement or in some cases, highlight underlying misclassification patterns.

The examples below include the sentence and its respective ID, the predicted label, the true label, and whether the prediction was correct.

**Table 28.**  
*Sentiment Analysis Predictions*

| ID      | Sentence                                                                                                                                                                                                                               | Prediction | True Label | Status    |
|---------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|------------|-----------|
| CIMB214 | Banks' net interest margins are likely to continue to be pressured by the two 25 bps rate cuts in late 2019, and two possible 25 bps policy rate cuts in 1H20, bringing the policy rate to 0.75%, below its historical floor of 1.25%. | Negative   | Negative   | Correct   |
| MB99    | Maybank is moving to harness such opportunities, including participation in the issuance of green securities in the ASEAN region.                                                                                                      | Neutral    | Positive   | Incorrect |
| BNM704  | The Bank further enhanced access to financial services, particularly to help households and micro businesses cope with financial challenges during the pandemic.                                                                       | Positive   | Positive   | Correct   |
| BNM316  | The growing availability of private healthcare treatment as an alternative to public hospitals has been accompanied by a rise in demand for medical and health insurance and takaful (MHI) in Malaysia.                                | Positive   | Neutral    | Incorrect |
| CIMB48  | These suppliers have a clearly articulated sustainability statement or a policy or have been supporting local communities and the                                                                                                      | Positive   | Positive   | Correct   |

|        |                                                                                                                                                                                                                                       |         |          |           |
|--------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|----------|-----------|
|        | environment through their CSR programmes and initiatives.                                                                                                                                                                             |         |          |           |
| BNM871 | During the year, the Bank continued to implement measures that are aimed at minimising disruption to our critical operations, ensuring staff safety and wellbeing, and managing vulnerabilities posed by hybrid working arrangements. | Neutral | Positive | Incorrect |

Table 28 illustrates six representative examples from the sentiment analysis test set, showing both successful and flawed predictions. The correct predictions (e.g., CIMB214, BNM704, CIMB48) indicate that the model is capable of distinguishing clear sentiment cues, such as negative outlooks tied to policy rate cuts or positively framed statements about access and sustainability. These results highlight the model’s strength in interpreting strongly worded financial implications or direct expressions of improvement or commitment.

However, incorrect predictions (e.g., MB99, BNM316, BNM871) show some of the model’s limitations in capturing nuanced sentiment. In the case of MB99, the mention of “harnessing opportunities” and “green securities” suggests a forward-looking, positive sentiment, yet the model defaulted to neutral—likely due to the subtlety of the optimistic tone. Similarly, for BNM316, although the growing demand for health insurance can be interpreted as a positive signal for the industry, the model labelled the sentence as neutral, missing the implicit positivity. These misclassifications reveal a tendency for the model to struggle when sentiment is embedded in implications rather than stated directly. Overall, while the sentiment model performs reliably on more explicit statements, it is less sensitive to implicit or indirect sentiment shifts common in financial reporting.

**Table 29.**  
*ESG Classification Predictions*

| ID     | Sentence                                                                                                                                                                                                        | Prediction    | True Label    | Status    |
|--------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|---------------|-----------|
| BNM571 | As stipulated in the Staff Welfare Account (Medical Fund) Trust Directions 2006, the Medical Fund shall be administered by a Medical Fund Committee.                                                            | Social        | Social        | Correct   |
| BNM519 | In 2019, we had bilateral meetings with counterparts from Bank Indonesia and the Monetary Authority of Singapore.                                                                                               | Governance    | None          | Incorrect |
| MB133  | In line with the terms of our Group Procurement Manual, any form of child labour is prohibited within our value chain.                                                                                          | Governance    | Governance    | Correct   |
| MB83   | Maybank recognises the environmental impact associated with coal-fired power generation and we are committed to supporting the transition to greener power generation.                                          | Environmental | Environmental | Correct   |
| MB85   | The Group's current exposure to coal-fired power plants is primarily concentrated in ASEAN countries with relatively low electrification rate where long-term electricity security and reliability is paramount | Environmental | None          | Incorrect |

|        |                                                                                                                                                              |        |      |           |
|--------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|------|-----------|
| BNM325 | While this is a positive development and a sign of progress in the country's health outcomes, this development naturally contributes to higher medical costs | Social | None | Incorrect |
|--------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|------|-----------|

For the ESG test set, the highlighted representative sentences reflect the model's ability to correctly classify ESG themes as well as its tendency to over-predict ESG relevance in sentences, even if the sentences does not thematically contain ESG content. Correct predictions such as BNM571, MB133, MB83, show that the model was successful in identifying key ESG elements, namely those related to social welfare, governance policies, and environmental commitments. In contrast, the incorrect predictions (BNM519, MB85, BNM325) contained common misclassifications due to the model concluding that there was ESG content present based on surface-level cues, when the true label was in face 'None'. These results suggest that while the model has improved and can now identify clear ESG signals, it still struggles to classify sentences that are more contextually nuanced or fringe cases where the ESG theme may be implied rather than explicitly stated.

**Table 30.**  
*FLS Classification Predictions*

| ID     | Sentence                                                                                         | Prediction | True Label | Status  |
|--------|--------------------------------------------------------------------------------------------------|------------|------------|---------|
| BNM687 | These companies currently offer digital insurance, peer-to-peer takaful and aggregator services. | Not-FLS    | Not-FLS    | Correct |

|         |                                                                                                                                                                                                            |                     |                     |           |
|---------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|---------------------|-----------|
| BNM1039 | This would ultimately serve towards preserving macroeconomic stability and protecting the welfare of Malaysians.                                                                                           | Non-specific<br>FLS | Not-FLS             | Incorrect |
| MB82    | To further strengthen the Bank's commitment as a responsible finance provider, more RACs are expected to be progressively introduced for other industries, including the power and real estate industries. | Non-specific<br>FLS | Not-FLS             | Incorrect |
| CIMB109 | Ultimately, we want to strengthen our position as the trusted financial advisor to our customers, enabled by personal financial management via our digital channels.                                       | Non-specific<br>FLS | Not-FLS             | Incorrect |
| BNM1201 | We will also continue to advance our research work and strengthen internal competence and awareness to build a green culture within our own operations.                                                    | Non-specific<br>FLS | Non-specific<br>FLS | Correct   |
| CIMB108 | With RM3.0 billion earmarked for SLLs, this incentive is effective from 1 January 2020 to 31 December 2024.                                                                                                | Specific-FLS        | Specific-FLS        | Correct   |

Table 30 presents the representative examples for the FLS classification task. It can be observed from the correct predictions (e.g., BNM687, BNM1201, CIMB108) that the model is able to accurately recognise sentences that either contain no forward-looking content, describe general intentions, or specify future initiatives with timeframes or quantifiable commitments.

However, several misclassifications reveal consistent areas of confusion. Sentences such as BNM1039, MB82, and CIMB109 were classified as ‘Non-specific FLS’ despite the clear lack of forward-looking elements or any mention of concrete action. While these sentences carry an aspirational tone (“preserving stability”, “strengthen our position”), they clearly do not contain clearly outlined future steps, making the ‘Not-FLS’ label assigned in the true label a more appropriate class. This suggests the model may be overly influenced by forward-oriented language without distinguishing between vision and verifiable action.

These errors indicate a recurring challenge in differentiating aspirational tone from genuine FLS content, particularly when domain-specific expressions of intent blur the lines between descriptive commentary and actionable forecast. As such, although the model performs well when FLS content is structurally explicit (e.g., timelines, financial allocations), it remains vulnerable to over-predicting in the presence of suggestive, yet ultimately static, statements.

#### **5.5.4 Error Analysis**

An in-depth analysis of the prediction errors that occurred during the classification process was conducted to further understand the recurring challenges faced by the model to correctly classify sentences to their appropriate label. Some of these challenges include difficulties with implicit content, overgeneralisations, the ambiguous nature of sentences, as well as a tendency to over-predict certain classes.

During the sentiment analysis task, most of the errors observed were a result of misinterpreting the tone or true purpose of the sentence. As seen in Table 28, sentences that were inherently positive in nature but presented in a neutral or descriptively ambiguous tone (MB99, BNM871) were incorrectly assigned the ‘Neutral’ label, when the actual label was ‘Positive’. This suggests that the model tends to rely more on obvious sentiment cues instead of considering contextual nuances or underlying intents, particularly when dealing with sentences that are presented in a more formal or factual tone. In contrast, sentences that containing positive initiatives presented in a neutral tone were sometimes under-classified, reflecting a general limitation of the model to distinguish between face-based descriptions and emotionally weighted statements. This poses a significant challenge, especially for financial document analysis where information is mostly presented in a formal and measured tone.

For the ESG classification task, most of the errors that occurred involved sentences that were annotated as ‘None’ but incorrectly predicted by the model as ESG-relevant content. For example, as shown in Table 29, sentences that were predicted as ‘Governance’ or ‘Environmental’ like BNM519 and MB85 respectively, lacked explicit ESG themes. Both sentences were descriptive in nature and focused more on general activities. This indicates that model may rely on the presence of thematic keywords, like “bilateral meetings “ or “coal-fired” without fully understanding the context of the sentence in relation to the ESG framework. The model is also prone to classify false positive and over-attributing ESG relevance, especially when faced with formal, institutional language that may only loosely reference ESG themes. There is also the case of the overlapping nature ESG categories, where a sentence may fall into two or more classes, which leads to misclassifications when only one label is allowed per sentence.

The FLS classification task also highlighted the model’s tendency to over-predict the ‘Non-specific FLS’ class when exposed to future-oriented or goal-setting language. Several

sentences that were noted in Table 30 (BNM1039, MB82, CIMB109) were misclassified as *Non-specific FLS* despite lacking any actionable forecast or time-bound plans, which would be required to meet the criteria of a forward-looking statement. This suggests that the model is particularly sensitive to lexical indicators of intent, such as words like “will” or “aim to”, even in the absence of a clear plan or commitment, reflecting that although the model performs well on sentences with clear indications of future plans (CIMB108), it still struggles to classify label accurately when statements lack specific time indicators.

Across all three classifications tasks, one clear fact stands out – the model is highly susceptible to surface-level cues rather than deeper semantic understanding. In industries like finance, where subtle contextual signals often determine the meaning and intent of a sentence, this behaviour may prove problematic. Furthermore, another reason for the misclassification is due to the class imbalances present in the dataset as well as the limited training examples available for minority labels, even when class weighing was applied as a precautionary step to offset these effects. For example, the ‘Social’ class in ESG and ‘Specific-FLS’ class both had relatively few examples in the training data, thereby reducing the model's exposure to varied linguistic expressions of these categories.

#### **5.5.5 Limitations of the Evaluation Approach**

While obvious improvements were observed across all tasks, several limitations, relating mostly to the dataset composition, contextual granularity, and class distribution, in the methodology must be acknowledged to inform future efforts

Firstly, the analysis was conducted at the sentence level, which does not fully capture the contextual flow present in financial documents. Many documents and disclosures rely on prior or subsequent sentences for the full meaning to be understood. Isolating sentences may hinder accurate classification.

Second, although the dataset's focus on Malaysian financial reports aligns with the study's core objective, the training set – comprising of 2,147 annotated sentences – was relatively small. The limited availability of diverse sentences may have constrained the model's exposure to the full range of linguistic variability found in financial documents. This is especially evident when considering the model's difficulty to handle minority classes and ambiguous expressions, where a broader, more varied dataset might have provided the contextual basis needed for more accurate classifications. Additionally, the single label approach, particularly for ESG classification, may have oversimplified instances where two or more domains overlapped. This then further limits the model's ability to capture the nuanced and often integrated nature of ESG content.

Finally, class imbalances frequently persisted across all tasks, particularly for Social in ESG and Specific-FLS, where both classes were the least represented in their respective datasets. Although fine-tuning improved some metrics for these labels, skewed distributions likely led to biased classifications towards more dominant classes like Governance and Non-specific FLS –which had the largest sample sizes – thus limiting the model's sensitivity to edge cases.

#### **5.5.6 Real-World Implications for Financial Text Analysis**

The results of this study reveal several meaningful observations into the practical use of language models in financial analysis, particularly for financial text analysis of corporate reports. Fine-tuning the FinBERT model on a custom annotated dataset contributed greatly to the improvement of the model's performance in classifying sentiment, ESG-related content, and forward-looking statements – all of which are highly relevant areas in today's financial and regulatory environments.

Financial analysts and institutional investors are likely to benefit from an accurate sentiment classification model that can flag tones within a document, whether optimistic or cautionary. The model in this study has shown favourable results in identifying clear sentiment polarity, even in sentences that lean towards a factual tone or performance-related statements. Thus, it is clear that utilising the model would greatly ease the process of reviewing documents, allowing for quicker assessments of company outlooks and risk levels.

The ability to identify environmental, social, and governance related content in financial disclosures is particularly significant in today's financial landscape, where ESG considerations and activities can influence the investment decisions and shape public perception of a company. As the demand for ESG transparency grows across global markets, the presence of a model that can reliably categorise ESG statements can reduce the manual burdens of reviewing long and technical documents. However, due to lower precision and recall for minority labels, this model may not suffice in high-stakes ESG evaluation scenarios as-is. Incorporating multi-label classification techniques or training ensemble models could help improve performance in detecting overlapping ESG themes and subtle linguistic cues.

The forward-looking statement (FLS) classification task also has notable implications. Being able to automatically identify specific FLS content is highly useful for regulators, risk analysts, and financial auditors who monitor predictive statements for compliance or strategic planning. The model's strength in classifying Not-FLS content suggests it can serve as an effective first-pass filter to separate factual disclosures from potentially speculative language. Nonetheless, its weakness in identifying vague, non-specific FLS implies the need for further model refinement. Techniques such as hierarchical classification or adding contextual windows (e.g., paragraph-level input) could improve recognition of ambiguous or loosely phrased future projections.

Overall, while the fine-tuned FinBERT model shows promising results and solid performance for certain tasks, it still somewhat falls short in achieving consistent accuracy across all categories. Future improvements and targeted enhancements may help increase reliability, leading to the model's deployment in real-world financial analysis workflows. With further refinement, the model holds immense potential as a valuable classification tool for supporting decision-making processes in finance.

## **5.6 Discussion of Findings**

The findings of this study address the core research objective: to analyse and evaluate the outcomes of the FinBERT model in providing insights corporate financial reports in the context of sentiment analysis, ESG evaluation, and predictive analytics on financial texts. In each of the tasks, the results have consistently shown that fine-tuning the FinBERT model on a local, annotated dataset led to an increase performance. Evaluation metrics such as accuracy, F1-score, and precision demonstrated marked increases post fine-tuning, particularly for underrepresented or nuanced categories. One example is the notable improvements for the sentiment classification tasks when classifying 'Positive' and 'Neutral' classes. Similarly, for ESG classification, fine-tuning demonstrated significant increases in performance for the 'Social' class, which was also underrepresented in the dataset. The FLS task, while experiencing challenges in classifying some predictive statements, showed better sensitivity towards specific forecasting statement as a result of fine-tuning.

These findings validate the importance of domain-specific model adaptation and localisation in NLP models. The pre-trained FinBERT model, which was originally trained on general financial corpora had limited comprehension of linguistic and contextual features of Malaysian financial reports. After fine-tuning with industry-specific content, the model was

more equipped to differentiate between overlapping classes and handle complex, context-sensitive statements

The results of the fine-tuned ESG model in particular, which previously measured a notably low accuracy of 33% before fine-tuning, align with those reported in Xia et al. (2024), who similarly found that fine-tuning significantly enhanced the performance of the FinBERT model for ESG classification. In their study, the baseline performance of the model achieved an accuracy of 72% and through targeted tuning with ESG-specific data, the performance increased to 91%. These findings, in truth, highlight a broader trend within NLP, where the effectiveness of pretrained models greatly increases when adapted with a specific linguistic, cultural or contextual dataset.

However, it is important to note that while the model's performance in this study improved across the board, several limitations were also observed. Sentence-level classification, class imbalances and overlapping content continue to pose significant hurdles to the model's accuracy, especially for fringe cases. These issues highlight the need for more granular annotations schemes or document-level modelling in future research.

Ultimately, the conclusions drawn from this study support the idea that fine-tuning domain-specific language models with local data can greatly improve classification performance. They reinforce the viability of deploying transformer-based models in niche NLP contexts, specifically within financial markets that are constantly evolving like that of Malaysia's financial landscape. Future work may extend the results presented here after increasing the dataset, shifting to multi-label classification approaches (especially for ESG), and incorporating additional contextual information beyond just sentence level.

## **5.7 Chapter Summary**

The chapter presented the evaluation and analysis of the pretrained and fine-tuned FinBERT models across three financial classification tasks. The model's performance using standard evaluation metrics were discussed and key findings were highlighted through quantitative results and qualitative error analysis. The real-world implications and potential improvements were also discussed. The chapter concludes that with further improvements, the fine-tuned models hold great potential for practical applications in financial report analysis.

## Chapter 6: Conclusion and Future Work

### 6.1 Introduction

This chapter summarises and concludes the study conducted on the FinBERT model. The chapter revisits the objectives and research questions outlined in the beginning and evaluates whether they were achieved and to what extent, reflecting on key findings from both the testing and fine-tuning phases. This chapter also outlines any limitations observed through the course of this study and provides recommendations for future improvement.

### 6.2 Achievements

In this section, the objectives and research questions are evaluated to see whether they were successfully achieved or not. Below are the research questions previously outlined in **Chapter 1**:

**Table 31.**  
*Summary of Research Questions*

| Research No. | Research Question                                                                                                                                                                                |
|--------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>RQ1</b>   | How effective is FinBERT in analysing financial texts to extract sentiment, evaluating ESG factors, and generating predictive outcomes from corporate reports in the Malaysian financial sector? |
| <b>RQ2</b>   | How can FinBERT support improved financial decision-making for both institutions and individual investors?                                                                                       |
| <b>RQ3</b>   | How can FinBERT help bridge the gap between financial literacy and practical application for investors and the general public?                                                                   |

Following the research questions, the objectives of this study are presented in the table below:

**Table 32.**  
*Summary of Research Objectives*

| <b>Objective No.</b> | <b>Research Objective</b>                                                                                                                                                                                      |
|----------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>OB1</b>           | To review the fundamentals of LLMs and FinBERT, financial issues, and the role of AI in finance.                                                                                                               |
| <b>OB2</b>           | To design and develop an approach for utilising the fine-tuned FinBERT model to conduct sentiment analysis, ESG evaluation, and predictive analytics on financial texts.                                       |
| <b>OB3</b>           | To analyse and evaluate the outcomes of the FinBERT model in providing insights corporate financial reports in the context of sentiment analysis, ESG evaluation, and predictive analytics on financial texts. |

To answer **Objective 1 and Research Question 1**, the study reviewed key concepts related to LLM, FinBERT and financial technologies. The studies outlines in Chapter 2 provide the necessary foundation for understanding how pretrained models such as FinBERT operate and how AI is being integrated into financial services. Relevant literature was presented to explore the three classification tasks and their importance, thereby setting the context for the research.

#### 1. Model Development and Implementation

To meet **Objective 2 and Research Question 1**, the FinBERT model was fine-tuned on a custom Malaysian financial text dataset using the Hugging Face transformers library. The fine-tuning process, outlined in Chapter 4, involved training three separate models for sentiment analysis, ESG classification, and FLS classification. Each model was trained using a training set of annotated sentences extracted from annual reports. Several appropriate hyperparameters and early stopping techniques were applied to optimise

learning and the entire process was conducted in a Google Colab environment. Each model was trained for up to five epochs and the performance was tracked using validation accuracy and loss.

This implementation marked a key step in adapting FinBERT to local financial language and context, enhancing its ability to interpret region-specific financial narratives. The successful training of these models enabled further evaluation and user interfacing, thereby fulfilling the technical development aspects of the project.

## 2. Performance Evaluation and Analysis

To address **Objective 3 and Research Question 2**, the fine-tuned models were evaluated in Chapter 5 based on several key metrics: accuracy, precision, recall, and F1-score. Sentiment classification showed favourable results post fine-tuning with notable improvement over the baseline model. The ESG model showed significant gains post fine-tuning while the FLS model also showed some increase. A qualitative analysis was conducted to better understand the errors present and how the models were acting at the sentence level. These final results showed that the model has the potential to expose trends in corporate narratives, identify ESG content, and indicate forward-looking language in formal financial disclosures.

## 3. Practical Accessibility via Gradio Interface

**Objectives 2 and 3 as well as Research Question 3** are directly addressed through the development of a Gradio interface, as detailed in Section 4.5. As part of the study's aim to bridge the gap between technical outputs and public usability, the interface marks a step forward in making AI models step forward in making AI-powered financial analysis tools more accessible. Users will be able to use the web-based interface developed to input single

sentences or upload Excel files for automatic classification using the fine-tuned model to gather insights from financial documents easily.

### **6.3 Constraints**

While the study successfully achieved its objectives and answered its research questions, several constraints were identified along the way. These included limited computational resources, a relatively small and imbalanced labelled dataset, and the inherent complexity of interpreting nuanced financial language. Additionally, the Gradio interface was deployed as a prototype rather than a production-ready tool due to time and infrastructure limitations. These factors influenced the scope and generalisability of the model's performance.

### **6.4 Future Work and Recommendations**

Considering the results of the study and insight gained from the analysis conducted, there are several recommendations and areas for future work

Firstly, using a larger number of annotated financial sentences to expand upon the dataset and build a more balanced corpus, specifically with underrepresented classes (e.g., "Social" and "Specific FLS"), would allow for better generalisability of the models built.

Secondly, incorporating additional NLP techniques, such as Optical Character Recognition (OCR) would allow both text and images to be extracted from the reports, thereby broadening the range of usable data. This could be especially beneficial when financial disclosures include embedded tables, scanned statements, or image-based reports not accessible through standard text extraction.

Lastly, the interface can be deployed as a scalable, cloud-based application programming interface (API), allowing real-world integration into financial tools, dashboards, or educational platforms, thus extending the usability of the model beyond purely experimental research.

## **6.5 Chapter Summary**

In summary, this chapter provides an overview of achievements, constraints, and areas of improvement observed throughout the study. Through a structured evaluation process, the model was adapted to perform sentiment analysis, ESG classification, and FLS classification on Malaysian financial text. The results demonstrate that the model is capable, despite limitations such as class imbalance. The inclusion of a Gradio interface demonstrated the practical application of the model. The outcomes gained offer important insights for future improvements and wider adoption of AI-driven financial text analysis tools.

## **6.6 Project Conclusions**

The project concludes with the successful fine-tuning and deployment of the FinBERT model for financial text analysis, thereby fulfilling the objectives set out in the early phases of the research. The study showed the model's potential to support classification tasks in domain-specific environments, where contextual and linguistic nuances play a critical role in interpreting financial narratives accurately. Through systematic evaluation and the deployment of the fine-tuned models via a Gradio interface, the study also bridged the gap between technical implementation and user accessibility. Despite the limitations present through the study, the findings highlight FinBERT's applicability for industry use and opens doors for future enhancements in AI-powered financial analysis tools.

## References

- Adesina, A. A., Iyelolu, T. V., & Paul, P. O. (2024). Leveraging predictive analytics for strategic decision-making: Enhancing business performance through data-driven insights. *World Journal of Advanced Research and Reviews*, 22(3), 1927–1934. <https://doi.org/10.30574/wjarr.2024.22.3.1961>
- Ahmad, H., Yaqub, M., & Lee, S. H. (2024). Environmental-, social-, and governance-related factors for business investment and sustainability: a scientometric review of global trends. In *Environment, Development and Sustainability* (Vol. 26, Issue 2, pp. 2965–2987). Springer Science and Business Media B.V. <https://doi.org/10.1007/s10668-023-02921-x>
- Araci, D. (2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. <http://arxiv.org/abs/1908.10063>
- Bahoo, S., Cucculelli, M., Goga, X., & Mondolo, J. (2024). Artificial intelligence in Finance: a comprehensive review through bibliometric and content analysis. *SN Business & Economics*, 4(2). <https://doi.org/10.1007/s43546-023-00618-x>
- Beltagy, I., Lo, K., & Cohan, A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text*. <http://arxiv.org/abs/1903.10676>
- Bialowolski, P., Cwynar, A., & Weziak-Bialowolska, D. (2022). The role of financial literacy for financial resilience in middle-age and older adulthood. *International Journal of Bank Marketing*, 40(7), 1718–1748. <https://doi.org/10.1108/IJBM-10-2021-0453>
- Cao, X., Li, S., Katsikis, V., Khan, A. T., He, H., Liu, Z., Zhang, L., & Peng, C. (2024). Empowering financial futures: Large language models in the modern financial landscape. *EAI Endorsed Transactions on AI and Robotics*, 3. <https://doi.org/10.4108/airo.6117>
- Çera, G., Khan, K. A., Mlouk, A., & Brabenec, T. (2021). Improving financial capability: the mediating role of financial behaviour. *Economic Research-Ekonomska Istrazivanja*, 34(1), 1265–1282. <https://doi.org/10.1080/1331677X.2020.1820362>
- Chen, C.-C., Huang, H.-H., & Chen, H.-H. (2019). Numeral Attachment with Auxiliary Tasks. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1161–1164. <https://doi.org/10.1145/3331184.3331361>
- Chen, C.-C., Huang, H.-H., Takamura, H., & Chen, H.-H. (2019). *Final Report of the NTCIR-14 FinNum Task: Challenges and Current Status of Fine-Grained Numeral Understanding in Financial Social Media Data* (pp. 183–192). [https://doi.org/10.1007/978-3-030-36805-0\\_14](https://doi.org/10.1007/978-3-030-36805-0_14)
- Chen, C.-C., Huang, H.-H., Takamura, H., & Chen, H.-H. (2020). *Overview of the NTCIR-15 FinNum-2 Task: Numeral Attachment in Financial Tweets*. [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html)

- Chopra, S. S., Senadheera, S. S., Dissanayake, P. D., Withana, P. A., Chib, R., Rhee, J. H., & Ok, Y. S. (2024). Navigating the Challenges of Environmental, Social, and Governance (ESG) Reporting: The Path to Broader Sustainable Development. In *Sustainability (Switzerland)* (Vol. 16, Issue 2). Multidisciplinary Digital Publishing Institute (MDPI). <https://doi.org/10.3390/su16020606>
- Chu, Z., Guo, H., Zhou, X., Wang, Y., Yu, F., Chen, H., Xu, W., Lu, X., Cui, Q., Li, L., Zhou, J., & Li, S. (2023). *Data-Centric Financial Large Language Models*. <http://arxiv.org/abs/2310.17784>
- Daudert, T., Buitelaar, P., & Negi, S. (2018). *Leveraging News Sentiment to Improve Microblog Sentiment Classification in the Financial Domain*. <https://www.statista.com>
- Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. <https://github.com/tensorflow/tensor2tensor>
- Du, K., Xing, F., Mao, R., & Cambria, E. (2024). Financial Sentiment Analysis: Techniques and Applications. *ACM Computing Surveys*, 56(9). <https://doi.org/10.1145/3649451>
- Financial Capability. (2021, November 15). World Bank Group. <https://responsiblefinance.worldbank.org/en/responsible-finance/financial-capability>
- Garg, N., & Singh, S. (2018). Financial literacy among youth. In *International Journal of Social Economics* (Vol. 45, Issue 1, pp. 173–186). Emerald Group Publishing Ltd. <https://doi.org/10.1108/IJSE-11-2016-0303>
- Glodd, A., Berlin, H., & Hristova, D. (2023). *Extraction of Forward-looking Financial Information for Stock Price Prediction from Annual Reports Using NLP Techniques*. <https://hdl.handle.net/10125/103313>
- Haque, S., Eberhart, Z., Bansal, A., & McMillan, C. (2022). Semantic Similarity Metrics for Evaluating Source Code Summarization. *IEEE International Conference on Program Comprehension, 2022-March*, 36–47. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>
- Head, C. B., Jasper, P., McConnachie, M., Raftree, L., & Higdon, G. (2023). Large language model applications for evaluation: Opportunities and ethical implications. *New Directions for Evaluation*, 2023(178–179), 33–46. <https://doi.org/https://doi.org/10.1002/ev.20556>
- Huang, A. H., Wang, H., & Yang, Y. (2022). *FinBERT-A Large Language Model for Extracting Information from Financial Text*.
- Kempson, E., Perotti, V., & Scott, K. (2013). *Measuring financial capability: a new instrument and results from low-and middle-income countries*. [www.worldbank.org](http://www.worldbank.org)
- Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3), 3713–3744. <https://doi.org/10.1007/s11042-022-13428-4>

- Kruse, L., Wunderlich, N., & Beck, R. (n.d.). *Artificial Intelligence for the Financial Services Industry: What Challenges Organizations to Succeed*.  
<https://hdl.handle.net/10125/60075>
- Lak, P., & Turetken, O. (2017). The Impact of Sentiment Analysis Output on Decision Outcomes: The Impact of Sentiment Analysis Output on Decision Outcomes: An Empirical Evaluation An Empirical Evaluation. In *AIS Transactions on Human-Computer Interaction*. [www.ratemds.com](http://www.ratemds.com)
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*.  
<http://arxiv.org/abs/1909.11942>
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>
- Lee, S. P., & Isa, M. (2023). Environmental, social and governance (ESG) practices and financial performance of Shariah-compliant companies in Malaysia. *Journal of Islamic Accounting and Business Research*, 14(2), 295–314. <https://doi.org/10.1108/JIABR-06-2020-0183>
- Li, F., Berger, P., Core, J., Dichev, I., Indjejikian, R., Lang, M., Leheavy, R., Lundholm, R., Nagar, V., Plumlee, M., Shakespeare, C., Shevlin, T., Skinner, D., Chen, M., Fang, Q., Li, Q., Liong, C., Lu, D., McMahon, A., ... Rencheng Wang, especially. (2008). *The Determinants and Information Content of the Forward-looking Statements in Corporate Filings - A Naive Bayesian Machine Learning Approach*.  
<http://dx.doi.org/10.2139/ssrn.1267235>
- Li, X., Chan, S., Zhu, X., Pei, Y., Ma, Z., Liu, X., & Shah, S. (2023). *Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks*. <http://arxiv.org/abs/2305.05862>
- Li, Y., Wang, S., Ding, H., & Chen, H. (2023). Large Language Models in Finance: A Survey. *ICAIF 2023 - 4th ACM International Conference on AI in Finance*, 374–382. <https://doi.org/10.1145/3604237.3626869>
- Li, Y., Yi, J., Chen, H., & Peng, D. (2021). Theory and application of artificial intelligence in financial industry. *Data Science in Finance and Economics*, 1(2), 96–116. <https://doi.org/10.3934/DSFE.2021006>
- Liang, Y., Liu, Y., Zhang, B., Wang, C. D., & Yang, H. (2024). *FinGPT: Enhancing Sentiment-Based Stock Movement Prediction with Dissemination-Aware and Context-Enriched LLMs*. <http://arxiv.org/abs/2412.10823>
- Liu, X.-Y., Wang, G., Yang, H., & Zha, D. (2023). *FinGPT: Democratizing Internet-scale Data for Financial Large Language Models*. <http://arxiv.org/abs/2307.10485>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. <http://arxiv.org/abs/1907.11692>

- Loginova, E., Tsang, W. K., van Heijningen, G., Kerkhove, L. P., & Benoit, D. F. (2024). Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data. *Machine Learning*, 113(7), 4761–4784. <https://doi.org/10.1007/s10994-021-06095-3>
- Loughran, T., & McDonald, B. (2020). Textual Analysis in Finance. *Annual Review of Financial Economics*, 12(Volume 12, 2020), 357–375. <https://doi.org/https://doi.org/10.1146/annurev-financial-012820-032249>
- Lui, T. K., Zainuldin, M. H., Wahidudin, A. N., & Foo, C. C. (2020). Corporate social responsibility disclosures (CSRDs) in the banking industry: a study of conventional banks and Islamic banks in Malaysia. *International Journal of Bank Marketing*, 39(4), 541–570. <https://doi.org/10.1108/IJBM-04-2020-0192>
- Malo, P., Sinha, A., Takala, P., Korhonen, P., & Wallenius, J. (2013). *Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts*. <http://arxiv.org/abs/1307.5336>
- Mariko, D., Labidurie, E., Ozturk, Y., Akl, H. A., & de Mazancourt, H. (2020). *Data Processing and Annotation Schemes for FinCausal Shared Task*. <http://arxiv.org/abs/2012.02498>
- Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. *IEEE Access*, 8, 131662–131682. <https://doi.org/10.1109/ACCESS.2020.3009626>
- Mokhtar, N., Sabri, M. F., & Ho, C. S. F. (2020). Financial Capability and Differences in Age and Ethnicity. *Journal of Asian Finance, Economics and Business*, 7(10), 1081–1091. <https://doi.org/10.13106/jafeb.2020.vol7.no10.1081>
- Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). *A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges*. <http://arxiv.org/abs/2406.11903>
- Omar, M., Brin, D., Glicksberg, B., & Klang, E. (2024). Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review. *American Journal of Infection Control*, 52(9), 992–1001. <https://doi.org/10.1016/j.ajic.2024.03.016>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825–2830.
- Peng, B., Chersoni, E., Hsu, Y.-Y., & Huang, C.-R. (2021). Is Domain Adaptation Worth Your Investment? Comparing BERT and FinBERT on Financial Tasks. In *Punta Cana and online*. <https://www.sec.gov/edgar.shtml>.
- Piotrowska, M. (2019). Facets of competitiveness in improving the professional skills. *Journal of Competitiveness*, 11(2), 95–112. <https://doi.org/10.7441/joc.2019.02.07>
- Potrich, A. C. G., Vieira, K. M., & Kirch, G. (2015). Determinants of financial literacy: Analysis of the influence of socioeconomic and demographic variables. *Revista*

*Contabilidade e Financas*, 26(69), 362–377. <https://doi.org/10.1590/1808-057x201501040>

- Rane, N. L., Rane, J., & Paramesha, M. (2024). Artificial Intelligence and business intelligence to enhance Environmental, Social, and Governance (ESG) strategies: Internet of things, machine learning, and big data analytics in financial services and investment sectors. In *Trustworthy Artificial Intelligence in Industry and Society*. Deep Science Publishing. [https://doi.org/10.70593/978-81-981367-4-9\\_3](https://doi.org/10.70593/978-81-981367-4-9_3)
- Ringgit Plus Malaysian Financial Literacy Survey. (2024). RinggitPlus Malaysia. <https://ringgitplus.com/en/>
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. <http://arxiv.org/abs/1910.01108>
- Santini, F. D. O., Ladeira, W. J., Mette, F. M. B., & Ponchio, M. C. (2019). The antecedents and consequences of financial literacy: a meta-analysis. *International Journal of Bank Marketing*, 37(6), 1462–1479. <https://doi.org/10.1108/IJBM-10-2018-0281>
- Scao, T. Le, Fan, A., Akiki, C., Pavlick, E., Ili’c, S., Hesslow, D., Castagn’e, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., ... Wolf, T. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *ArXiv*, [abs/2211.05100](https://arxiv.org/abs/2211.05100). <https://api.semanticscholar.org/CorpusID:253420279>
- Shakil, M. H., Mahmood, N., Tasnia, M., & Munim, Z. H. (2019). Do environmental, social and governance performance affect the financial performance of banks? A cross-country study of emerging market banks. *Management of Environmental Quality: An International Journal*, 30(6), 1331–1344. <https://doi.org/10.1108/MEQ-08-2018-0155>
- Shen, Y., & Zhang, P. K. (2024). *Financial Sentiment Analysis on News and Reports Using Large Language Models and FinBERT*.
- Sipola, J., Saunila, M., & Ukko, J. (2023). Adopting artificial intelligence in sustainable business. *Journal of Cleaner Production*, 426. <https://doi.org/10.1016/j.jclepro.2023.139197>
- Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. (2018). Big Data: Deep Learning for financial sentiment analysis. *Journal of Big Data*, 5(1). <https://doi.org/10.1186/s40537-017-0111-6>
- Taskin, B., & Akal, F. (2024). Tales of Turbulence: BERT-based Multimodal Analysis of FED Communication Dynamics Amidst COVID-19 Through FOMC Minutes. *Computational Economics*. <https://doi.org/10.1007/s10614-023-10533-w>
- Wang, L., Cheng, Y., Xiang, A., Zhang, J., & Yang, H. (2024). *Application of Natural Language Processing in Financial Risk Detection*. <https://doi.org/10.48550/arXiv.2406.09765>
- Webersinke, N. (2023). *Natural Language Processing meets Accounting and Finance: Review and Performance Comparison of Textual Analysis Approaches*. <https://ssrn.com/abstract=4527724>

- World Bank Group. (2021, November 15). *Financial Capability*.  
<https://responsiblefinance.worldbank.org/en/responsible-finance/financial-capability>
- Wu, S., Irsoy, O., Lu, S., Dabrovolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). *BloombergGPT: A Large Language Model for Finance*. <http://arxiv.org/abs/2303.17564>
- Xia, L., Yang, M., & Liu, Q. (2024). *Using Pre-trained Language Model for Accurate ESG Prediction*. <https://huggingface.co/>
- Xing, F. Z., Malandri, L., Zhang, Y., & Cambria, E. (2020). *Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets*. Online.  
<https://sentic.net/stocksen.zip>
- Xu, H. (2023). *No Train Still Gain. Unleash Mathematical Reasoning of Large Language Models with Monte Carlo Tree Search Guided by Energy Function*.  
<http://arxiv.org/abs/2309.03224>
- Xu, H., Gan, W., Qi, Z., Wu, J., & Yu, P. S. (2024). *Large Language Models for Education: A Survey*. <http://arxiv.org/abs/2405.13001>
- Yang, H., Liu, X.-Y., & Wang, C. D. (2023). *FinGPT: Open-Source Financial Large Language Models*. <http://arxiv.org/abs/2306.06031>
- Yang, Y., UY, M. C. S., & Huang, A. (2020). *FinBERT: A Pretrained Language Model for Financial Communications*. <http://arxiv.org/abs/2006.08097>
- Zhao, H., Liu, Z., Wu, Z., Li, Y., Yang, T., Shu, P., Xu, S., Dai, H., Zhao, L., Mai, G., Liu, N., & Liu, T. (2024). *Revolutionizing Finance with LLMs: An Overview of Applications and Insights*. <http://arxiv.org/abs/2401.11641>

# Appendix A: Annotator Appointment Letter

**Universiti Malaysia Sarawak (UNIMAS)**  
Faculty of Computer Science and Information Technology  
Jln Datuk Mohammad Musa,  
94300, Kota Samarahan,  
Sarawak, Malaysia  
81780@siswa.unimas.my  
+6012 7010 754  
16 March 2025

---

**Philip Wong**  
17, Jalan PJU 1A/26E,  
Ara Damansara,  
47301 Petaling Jaya,  
Selangor

3 April 2025

Dear Mr. Wong,

Appointment as Annotator for Final Year Project Dataset

This letter serves as a formal invitation to participate as an annotator for my Final Year Project (FYP) at Universiti Malaysia Sarawak. The project, titled "*A Study of FinBERT's Effectiveness in Financial Text Analysis: Insights into Sentiment, ESG Factors, and Predictive Outcomes from Malaysian Financial Institutions*" (2025), involves the annotation of a dataset compiled from annual reports of selected local banks. The objective is to evaluate the effectiveness of the FinBERT model in conducting sentiment analysis, ESG (Environmental, Social, and Governance) classification, and forward-looking statement (FLS) classification.

As an experienced professional in finance and economics, expertise in the subject matter is invaluable in ensuring the accuracy and reliability of the dataset annotations. The role will involve reviewing and labelling text data based on a predefined criteria, contributing to the development of a comprehensive corpus.

The appointment as an annotator will commence from 5 April 2025 and is expected to conclude 20 April 2025. The annotation process will be facilitated through Microsoft Excel, and detailed guidelines will be provided to ensure consistency and clarity.

Contributions will be duly acknowledged in the research work, and sincere appreciation is extended for the support in advancing this study.

Thank you for considering this request.

Sincerely,  
Hanna Ellisa Joseph  
Final Year Student,  
Bachelor of Computer Science with Honours  
(Information Systems)  
Universiti Malaysia Sarawak (UNIMAS)

Ts. Dr. Sarah Flora Samson Juan,  
Supervisor  
Faculty of Computer Science and Information Technology  
(FCSIT)

## Appendix B: Annotator Form

### Annotator Form

#### SECTION 1: PERSONAL INFORMATION

Full Name: *Philip Wong Wan Ming*

Email Address: *phiwong60@gmail.com*

Contact Number: *012-2185891*

---

#### SECTION 2: PROFESSIONAL BACKGROUND

Field of Expertise: *Human Resource Management*

Years of Experience: *30+*

Prior Experience in Data Annotation (~~Yes~~/No):

If Yes, please provide details: \_\_\_\_\_

---

#### SECTION 3: ANNOTATION PROCESS ACKNOWLEDGMENT

Availability for Annotation Task:

Start Date: *16 April 2025*

End Date: *20 April 2025*

Preferred Mode of Communication (~~Email~~/Phone/Others):

Agreement to Follow Annotation Guidelines (Yes/~~No~~):

---

#### SECTION 4: DECLARATION

I, Philip Wong Wan Ming, hereby confirm my willingness to participate as an annotator for the Final Year Project at Universiti Malaysia Sarawak. I understand the requirements of the annotation task and agree to adhere to the guidelines provided.

Signature: 

Date: *16 April 2025*