A blue and red text on a black background

Description automatically generated

WIH3001 DATA SCIENCE PROJECT

PROJECT PROPOSAL

PROJECT TITLE: ANALYSE AND SUMMARISE DASHBOARD IMAGES USING LLM MODEL

SUPERVISOR: A.P. DR. MAIZATUL AKMAR BINTI ISMAIL

MATRIC NUMBER: S2114212

NAME: LIM HON TING

# Table of Content

[1) TITLE OF PROJECT 2](#_Toc149259122)

[2) INTRODUCTION/ BACKGROUND 2](#_Toc149259123)

[3) PROBLEM STATEMENT 3](#_Toc149259124)

[4) OBJECTIVES 3](#_Toc149259125)

[5) DATA SCIENCE METHODOLOGY 4](#_Toc149259126)

[6) REFERENCES 6](#_Toc149259127)

# TITLE OF PROJECT

The title proposed for this project is analyse and summarise dashboard images using LLM model.

# INTRODUCTION/ BACKGROUND

In various industries, valuable data for analysis is often spread across an assortment of financial reports, each stored in separate PDF files. Manually aggregating and extracting data points from this multitude of documents is a labour-intensive and time-consuming process (Myatt and Johnson, 2009). Significant demands on human resources required to collect data from multiple sources as well as the risk of errors and inconsistencies elevates in the analysis.

This is where data visualization emerges as a crucial tool for streamlining and optimizing the analysis of such dispersed financial data. Data visualization techniques are powerful tools that transform complex and scattered financial data into clear and understandable graphical representations. Data visualization simplifies the process of data interpretation by translating financial metrics, trends, and performance indicators into visual charts, graphs, and dashboards (Suprata, 2019). Decision-makers across various sectors can swiftly and intuitively access critical insights, make data comparisons across different reports, and unveil patterns that might otherwise remain concealed within the dense text of documents (Adhikari, 2015).

Moreover, the remarkable progressions in large language models (LLMs) towards the end of 2022 have opened new horizons in automating the interpretation and distillation of information from dynamic dashboard representations. LLMs now provide a transformative solution for this endeavour. Their ability to comprehend and summarize complex data makes them an indispensable tool in streamlining the analysis process (Huang et al., 2023). By effectively utilizing the comprehensive and varied data displayed in dashboards, LLMs are able to rapidly and precisely extract key findings, analyse data across various reports, and reveal hidden patterns in the visual representations. This automated approach boosts the efficiency of the analysis while significantly decreasing the chance of human biases affecting decision-making. The most critical insights can be presented in a user-friendly format, enabling decision-makers to grasp them quickly.

With recent advancements in technology, this project focuses on utilizing the power of LLMs to automate the analysis and summarization of findings from dashboards, addressing the persistent challenge of scattered data within the utilities sector. By adopting this approach, decision-makers can gain valuable insights, data-driven processes can become more efficient, and ultimately, both the industry and the public can reap the benefits.

# PROBLEM STATEMENT

In the current data-driven era, organizations across various sectors heavily rely on data visualization through dashboard images to inform critical decision-making processes. These dashboards play a crucial role in monitoring performance, tracking key metrics, and driving strategic initiatives and it is filled with charts, graphs, and textual information (Sadiku et al., 2016). However, extracting actionable insights from these visual data representations often require thorough manual analysis, leading to inefficiencies, delays, and the likelihood of human bias. Needless to say, human analysts bring their own perspectives and biases to the analysis process. These biases will affect how the data to be interpreted and conclusion drawn based on the data. Errors can occur and insights can become less accurate due to inconsistent interpretations and the possibility of subjectivity in analysis.

The current practices often require significant human involvement and domain expertise, making the process time-consuming and may contain subjective interpretations (Andrienko et al., 2020). Furthermore, as organizations generate vast amounts of data and increasingly rely on visual representations, the challenge of deriving meaningful conclusions becomes more pronounced because analysts often need to look at each element of a dashboard, understand its context and draw meaningful conclusions. This problem is of profound significance in this domain where timely decision-making is paramount.

To put it concisely, this project aims to address the following problems:

1. Manual analysis is time-consuming and labour-intensive, and there is a chance of human biases.
2. Inefficiencies in extracting insights, straining resources and slowing down the decision-making process.
3. Delay in generating insights from the data can cause the data out of date by the time analyst has done the analysis.

# OBJECTIVES

There are four objectives proposed for this project which are:

* To identify existing work on dashboard analysis using LLM model.
* To train a model that able to summarise the findings from the dashboard through LLM model.
* To develop an automated system that is capable of analyse dashboard and free from subjective interpretations.
* To evaluate the trained model using Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics.

# DATA SCIENCE METHODOLOGY

Diagram of a diagram of data

Description automatically generated

Figure 1 CRISP-DM (Huber et al., 2019)

* Business Understanding

Business understanding is the initial phase of CRISP-DM model. A preliminary project is developed based on our understanding in the project objectives and requirements from a business perspective.

* Data Understanding

In this phase, data acquisition and Exploratory Data Analysis (EDA) will be performed to discover insights from the dataset. The data that will be used in this project will be collected from the financial report of utilities companies.

The main point of doing EDA is to perform initial investigations on data to test our hypothesis on the dataset and verify our assumptions through summary of statistics and graphical representation (Patil, 2022). In this project, EDA will be used to understand the relationship between financial terms in that particular financial year.

* Data Preparation

Data preparation is the process which may take up to 80% of the time spent in the project. It involves collecting, cleaning, and labelling raw data and convert it into a form which is suitable for machine learning algorithms (What Is Data Preparation? - Data Preparation Explained - AWS, n.d.). In this project, a dashboard will be created by using the data which is obtained from the financial report.

* Modelling

Various modelling techniques will be tried out in this phase, and the parameters for the algorithms will be calibrated to execute the data mining tasks on pre-processed data. Since the data will be in dashboard, it is crucial to transform all the important information from the dashboard into textual data.

The dashboard will be divided into different parts based on the chart it has for the modelling. The first stage will be identifying the type of the chart and split them into different model for the next stage. The second stage will be data extraction from the chart by calculating the data range and pixel range from the image to map the data point to the actual value. Model that will be employed includes but not limited to Corner Net with Hourglass Net as backbone structure. At the end, the data from the model will be used to feed the Large Language Model (LLM) to analyse and summarise the finding.

* Evaluation

Evaluation metrics will be mainly focus on accuracy of the model such as RMSE and R2 to identify how accurate of the model on predicting the company performance in the future. Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2001) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) will be used for evaluating the quality of summaries generated by LLMs by compare the generated text with the reference text.

* Deployment

Once the best performing model is built and refined, it will be incorporated into a web dashboard by integrating it into the pipeline. The open-source app framework Streamlit is utilized to build the web dashboard. By uploading a dashboard image on the web dashboard, users can simulate the entire inference process.

In addition, the project's background, problem statements, and objectives, as well as visualization of the findings in the EDA process about this project can be found inside the web dashboard.

# REFERENCES

Myatt, G. J., & Johnson, W. P. (2009). *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications*. John Wiley & Sons.

Suprata, F. (2021). Data Storytelling with Dashboard: Accelerating Understanding Through Data Visualization in Financial Technology Company Case Study. *Metris: Jurnal Sains Dan Teknologi*, *20*(1). https://doi.org/10.25170/metris.v20i1.2375

Adhikari, S. (2015, May 21). *Developing dashboards showing financial key perfomance indicators: Case: City of Vantaa*. Theseus. https://urn.fi/URN:NBN:fi:amk-201505117265

Huang, A., Wang, H., & Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research*, *40*(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Sadiku, M., Shadare, A. E., Musa, S. M., Akujuobi, C. M., & Perry, R. (2016). Data visualization. *International Journal of Engineering Research and Advanced Technology (IJERAT)*, 2(12), 11-16. https://ijerat.com/index.php/ijerat/article/view/191

Andrienko, G., Andrienko, N., Drucker, S., Fekete, J. D., Fisher, D., Idreos, S., ... & Sharaf, M. (2020, March). Big data visualization and analytics: Future research challenges and emerging applications. In *BigVis 2020-3rd International Workshop on Big Data Visual Exploration and Analytics*. https://hdl.handle.net/1721.1/132286

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. *Procedia Cirp*, *79*, 403-408. https://doi.org/10.1016/j.procir.2019.02.106

Patil, P. (2022, May 30). What is Exploratory Data Analysis? - Towards Data Science. Medium. https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

*What is Data Preparation? - Data Preparation Explained - AWS*. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/what-is/data-preparation/#:~:text=Data%20preparation%20is%20the%20process,exploring%20and%20visualizing%20the%20data.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July).Bleu: a method for automatic evaluation of machine translation*. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (*pp. 311-318*).* https://doi.org/10.3115/1073083.1073135

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. *In* *Text summarization branches out* (pp. 74-81). https://www.semanticscholar.org/paper/ROUGE%3A-A-Package-for-Automatic-Evaluation-of-Lin/60b05f32c32519a809f21642ef1eb3eaf3848008