Unveiling Patterns in Online English Business News in Sri Lanka: Abstractive Text Summarization with Large Language Models

A Thesis Submitted for the Degree of Master of Business Analytics

V.M. Wimalarathna
University of Colombo School of Computing
2025

DECLARATION

Name of the Student: V.M. Wimalarathna
Registration Number: 2022/BA/027
Name of the Degree Programme: Master of Business Analytics
Project/Thesis Title: Unveiling Patterns in Online English Business News in Sri Lanka: Abstractive Text Summarization with Large Language Models

1. The project/thesis is my original work and has not been submitted previously for a degree at this or any other University/Institute. To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.

2. I understand what plagiarism is, the various types of plagiarism, how to avoid it, what my resources are, who can help me if I am unsure about a research or plagiarism issue, as well as what the consequences are at University of Colombo School of Computing (UCSC) for plagiarism.

3. I understand that ignorance is not an excuse for plagiarism and that I am responsible for clarifying, asking questions and utilizing all available resources in order to educate myself and prevent myself from plagiarizing.

4. I am also aware of the dangers of using online plagiarism checkers and sites that offer essays for sale. I understand that if I use these resources, I am solely responsible for the consequences of my actions.

5. I assure that any work I submit with my name on it will reflect my own ideas and effort. I will properly cite all material that is not my own.

6. I understand that there is no acceptable excuse for committing plagiarism and that doing so is a violation of the Student Code of Conduct.

Signature of the Student	**Date (DD/MM/YYYY)**
	15-Jun-2025

Certified by Supervisor(s)

This is to certify that this project/thesis is based on the work of the above-mentioned student under my/our supervision. The thesis has been prepared according to the format stipulated and is of an acceptable standard.

	Supervisor
Name	Dr. Randil Pushpananda
Signature	Randil
Date	15th June 2025

I would like to dedicate this thesis to all those who supported and encouraged me throughout this journey.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my research supervisor, Dr. Randil Pushpananda, for his continuous support, invaluable guidance, and encouragement throughout this research. His expertise and insights have greatly influenced the development of my work.

I am sincerely thankful to Mr. Rangana Jayashanka for his assistance and guidance in coordinating research activities and providing essential support during the research process.

My heartfelt thanks also go to Dr. Kasun Karunanayaka for his constructive feedback, which has significantly contributed to improve the quality of my research project.

Finally, I would like to extend my deepest appreciation to my family and friends for their constant support, encouragement, and belief in me. Your support has been a constant source of strength.

Thank you to everyone who has played a part in this journey.

ABSTRACT

The emergence of online business news has amplified the rapid generation of large volumes of textual data, making it an important resource for businesses to extract business insights to support business decision making. The aim of this research project is to propose an approach to leverage online English business news to gain business insights with specific focus on Sri Lanka. Accordingly, the research employs a combined approach of document classification, document clustering, and abstractive text summarization methods to utilize large amount online business news data to gain meaningful insights.

The research methodology incorporates Bidirectional Encoder Representations from Transformers (BERT) model for news classification, followed by K-means news clustering, and cluster-wise abstractive text summarization utilizing large language models (LLMs), including Llama3, Mistral, and DeepSeek R1 along with retrieval-augmented generation (RAG) to enhance the accuracy and relevance of the generated content. The dataset is obtained from Sri Lankan business news sources, including DailyFT and Lanka Business Online. The ability to fine-tune BERT for specific classification tasks, combined with the scalability of K-means clustering and the text generation capabilities of LLMs accommodates unveil business insights from large volumes of textual data. However, there is still room for improvement by experimenting with large datasets and improving above techniques to enhance the performance of proposed approach.

The results of the study indicate that the combination of machine learning methods, such as classification and clustering, along with LLMs, yields better results by complementing each method's strengths, to leverage textual data more efficiently than applying them in isolation. This makes the proposed approach is more pragmatic way of leveraging computational method to handle large amount of online business news. Therefore, current study lays the foundation for harnessing the capabilities of machine learning and LLMs for real-world business applications, specifically in developing tools that support decision-making in business by utilizing online business news data.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ARI	Adjusted Rand Index
ATS	Automatic Document Summarization
BERT	Bidirectional Encoder Representations from Transformers
FT	Flat Transformer
GPT	Generative Pre-trained Transformer
GUI	Graphical User Interface
HT	Hierarchical Transformer
IR	Information Retrieval
IT	Information Technology
KNN	K-nearest Neighbors
LLMs	Large Language Models
LR	Logistic Regression
LSTM	Long Short Term Memory
ML	Machine Learning
NB	Naïve Bayes
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
NMF	Non-negative Matrix Factorization
RF	Random Forest
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
AUC-ROC	Receiver Operating Characteristic Curve
RAG	Retrieval Augmented Generation
SVD	Singular Value Decomposition
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
WCSS	Within Cluster Sum of Square

CHAPTER 1
INTRODUCTION

1.1 Project Overview

The advancement of information technology (IT) has been shaping all aspects of businesses while making the business environment challenging and competitive. Particularly, technology has become a key factor for the sustainability of any business. The technology helps sustain the businesses by transforming business processes into digital form. On the other hand, technology as a tool supports business sustainability by empowering business decision making. Decision making in business involves tasks such as strategy formulation, competitor analysis, new product launch etc. In the modern business environment, technology has become a critical factor by providing necessary information to make business decisions. Further, rise in big data has made the variety of available data for decision making in different forms such as text, images, audios, and videos. Hence, making use of these diverse types of data to make decisions emerges as a critical factor for the businesses.

Among the various types of data, textual data has gained considerable attention due to its ability to support gaining business insights by giving hidden information in text documents (Kopackova, Komárková and Sedlak, 2008). Business insights lay the foundation for making business decisions. In other words, the knowledge obtained through business insights helps businesses to make solid business decisions (Ismail, Chukwuka and Abiola, 2024). The online availability of textual data in higher volumes, and the possibility of containing rich information in textual data have made the textual data more important for obtaining business insights. However, the challenge is how to utilize large amount of textual data to support business decision making. The currently available research studies have made efforts to apply text analytics tools in the domains such as information retrieval (IR), and natural language processing (NLP) to address that problem. Nevertheless, the aim of this research is to improve those efforts by combining techniques to make more actionable insights from textual data.

Online news plays a critical role when it comes to textual data as an important mode for information dissemination. However, all online news data would not be important equally. Therefore, it is useful to select most relevant online news for a given purpose. This can be done by assigning text documents into one or more classes or categories (Sarkar, 2019). On the other hand, even among the most relevant news, identifying natural grouping of online news is more

useful rather than pursuing them directly. The retrieval of the news grouping can be done by utilizing some similarity measures (Sarkar, 2019). The grouped textual data enables businesses to obtain more sense of online news data when making decisions. To get further information about those groups, NLP techniques such as text summarization can be applied. Summarized information is useful to gain the gist of large amounts of textual data. Therefore, categorizing online news, identifying grouping of online news, and generation of text summaries for those groups are the focus areas of this research project.

1.2 Motivation

The motivation of the study can be divided into two components. First, the current studies available have suggested many approaches to overcome the issues of processing large amounts of online news data. However, still, there have been suggestions to utilize more diverse datasets (Shevendrakumar, 2023). Further, the current research studies have been mostly focusing only on either document classification methods, document clustering or text summarization methods such as extractive or abstractive summarization methods. However, to the best of our knowledge, an approach to classify online news followed by clustering online news and make meaningful summaries for those clusters employing abstractive summarization methods have not been explored. Thus, it opens a reasonable research gap to be explored.

Second, researcher's profession involves creating content for a business information platform, where company news research is part of the daily work of the researcher. Accordingly, the researcher gathers, refines, and maps online business news data daily. However, that process is not relying on much of the text analytics techniques. Hence, utilizing text analytics techniques on online business news is useful to uncover insights that can be helpful to enrich the content in business information platform. Therefore, there is a career motivation to improve that content as that is a key to progress in the researcher's career too. Further, during the master's degree programme, sufficient level of studying was done on data analytics techniques such as machine learning, and text analytics. Thus, obtaining hands-on experience in applying those techniques learned to solve real world problem is another motive to conduct this research study.

1.3 The Project Problem, Questions, and Objectives

1.3.1 Research Problem

As mentioned above, the online news is capable of empowering business decision making by providing actionable insights. However, for the businesses to gain meaningful insights, all types of online news may not be useful as mentioned previously. In fact, businesses may want to exclude some news categories that are of comparatively low significance, despite being related to businesses in a broader context. Hence, focusing on online "business news" while excluding news categories such as political, sports, and entertainment is more suitable for businesses. This ensures to select data that is suitable for a given purpose. Further, it makes sure to establish a clear scope for news selection too. After excluding irrelevant news articles, it would be more valuable to group the remaining relevant news articles on their similarity. That facilitates the extraction of business insights specific to each group, ensuring a comprehensive understanding from multiple perspectives. Each news group contains multiple news documents; hence, generating an abstractive summary for each group can further enhance the clarity and usability of the extracted insights by capturing the key information in a concise and coherent manner. Based on those requirements, the research problem of this study is to propose a combined method to utilize online business news to gain business insights.

Business insight is something that helps understanding market conditions, business trends which in turn support the decision making of a business. Therefore, gaining business insights are crucial for a business to make innovations, capitalize opportunities in the market, compete in the market, survival in the business for long time, ultimately gaining competitive advantages (Ismail, Chukwuka and Abiola, 2024). Based on the above research problem, the research aims to propose an approach to utilize online business news in Sri Lanka to gain meaningful business insights. Hence, the goal of the research including developing a methodology to deal with online business news related to Sri Lanka, so that business firms can employ that to gain business insights to make sound business decisions.

1.3.2 Research Questions

Based on the above research problem, specific research questions can be derived as follows.

RQ1. What methods can be utilized to classify online business news for decision-making relevance followed by clustering relevant online business news into meaningful groups?

RQ2. What methods can be utilized to generate abstractive text summary for a group of online business news to unveil business insights?

RQ3. How can the abstractive text summarization method be enhanced when generating summaries for group of online business news?

1.3.3 Research Objectives

To answer the above research questions, the aim of the research is converted into actionable terms to make the focus of the research clear. Accordingly, specific research objectives to address above three research questions as follows.

Objective 1: To utilize suitable document classification method followed by suitable clustering method to categorize and group online business news.

Objective 2: To utilize LLMs for abstractive text summarization for generating meaningful summary about each group of online business news.

Objective 3: To utilize RAG method to enrich the abstractive text summary generated by LLMs.

1.4 Background of the Study

A task of compressing a collection of textual data into a shorter version while keeping the crucial information and meaning is known as summarization (Yadav, Desai and Yadav, 2022). The importance of text summarization has gained considerable attention amidst the rising number of online news sources. Mainly, the increasing number of online news has made the human brain incapable of processing them all together. Accordingly, the efficient and effective

processing of large amounts of online textual data has become cumbersome. However, based on the purpose of the user, some news may be important, and some news may be less important. Therefore, a mechanism to include some news and exclude some is an important aspect to address. Accordingly, prior to performing summarization, selecting relevant news is more suitable way as it helps reduce the inclusion of unnecessary information in generated summaries. Thus, the requirement for methods to handle that needs to be explored.

To filter the news based on the requirement of the user, text document classification is a suitable mechanism. In text document classification, the effort is to automatically classify the text documents into one or more categories. In broader level, text document classification can be conducted by utilizing traditional machine learning techniques or neural network-based classification models. Studies of Ahmed and Ahmed (2021), and Daud *et al.* (2023) utilized conventional methods such as Naïve Bayes, support vector machine, k-nearest neighbors, and logistic regression for document classification. On the other hand, researchers such as Nugroho, Sukmadewa and Yudistira (2021), and Song (2022) utilized neural network-based text document classification methods such as BERT. In recent years, neural network-based classification methods have gained popularity due to their ability to eliminate the need for feature engineering. Furthermore, pre-trained models such as BERT do not require training from scratch and only needs to be fine-tuned on specific tasks instead. Therefore, it would be easier to utilize pre-trained neural network-based methods over traditional approaches for filtering news documents from large collections, particularly for specific applications such as business decision-making.

Based on the number of input documents, text summarization can be divided into single document text summarization, and multi-document text summarization. Among the two approaches, multi-document text summarization is considered a complex task as it faces the problem of data redundancy. Multiple documents may contain the same information, and the generated summaries could repeat that same information which results in data redundancy. However, the increasing number of online news highlights the importance of multi-document text summarization over single-document text summarization.

Further, depending on the method employed in summary generation, text summarization can be divided into extractive text summarization, and abstractive text summarization. Extractive methods, extract key sentences from a given document/s and present those as the summary. Here, new sentence is not generated (Moratanch and Chitrakala, 2016). On the other hand, abstractive methods generate new sentences by paraphrasing the ideas in the original

document/s. Meanwhile, despite the popularity of extractive methods in the past, the recent developments in LLMs have made the researchers focus on utilizing LLMs to generate abstractive text summaries (Basyal and Sanghvi, 2023). OpenAI's generative pre-trained transformer (GPT) models, Meta's LLaMA models, Google's PaLM models, Mistral AI, and the recent DeepSeek R1 are some examples of LLMs.

The generation of abstractive summary for multiple documents creates value by compressing large amounts of data into digestible form by human. However, this can be further improved by grouping the online news initially and generating summaries for each group. In other words, grouping the large amount of online news and generating summaries for each group would give more sense of the news rather than summary generation for each document individually or a single summary for whole document collection. Therefore, for a collection of multiple documents, grouping them based on similarity before summarization would be a more efficient approach to obtain a summary in a meaningful way and capturing granular-level details. For that document clustering method is an effective way. Studies such as Kumbhar *et al*. (2020), and Shevendrakumar (2023) have utilized document clustering methods to cluster online news documents.

1.5 Research Significance

This research contributes to the domain of text analytics in several ways. First, the dataset that is utilized in the research is a combination of several sources. This is one of the suggestions made by the previous studies to diversify the dataset (Shevendrakumar, 2023). Second, leveraging pre-trained models such as BERT, alongside other techniques, to filter news is a key contribution of this study, particularly for businesses to gain business insights. Third, to the best of the knowledge, most of the previous studies have focused only on either news classification, clustering online news or abstractive text summarization. However, the novelty of this study is the combining of all three techniques to gain business insights which in turn support business decision making. Further, previous studies have paid attention to generation summaries for individual news or for whole collection of news. However, this study attempts to generate summaries for each cluster of online news that is something has not been explored before. Overall, combining document classification, document clustering, and abstractive summarization is the novelty of this study.

It is important to note that the approach suggested by this study is beneficial for business firms as an effective and efficient method of dealing with large amount of online business news data.

Particularly, this can be utilized as a supportive tool to gain business insights and support the decision makings by automating the processing of utilizing online business news data. Further, it could help businesses to update their knowledge about the things happening in the business environment. Furthermore, the capability of adding more and more sources along with grouping them makes the proposed approach a scalable, versatile solution.

1.6 Scope of the Study

The scope of this study is identified under several criteria. Some are related to the section of dataset, and time considered for selecting data. On the other hand, the scope is also defined in terms of tasks performed during the research study. This clearly identifies what is addressed by the research and what is not, providing a clear path to proceed. The following can be identified as the criteria for defining the scope of this study.

- The dataset that is utilized in this study is online business news relevant to Sri Lanka. Hence, the geographical scope of this study is Sri Lanka.

- The time period considered for collecting online news is from 2019 to 2024. Hence, the news for years beyond that period are out of scope.

- The research intends to utilize open-source LLMs. Hence, building an LLM specifically for this task is not within the scope of this study.

1.7 Feasibility Study

The study mostly relies on open-source Python libraries. For the study, online news data is collected from online news websites such as DailyFT, and Lanka Business Online. These two sources allow free access to online news without any requirement of paying a subscription. For the collection of data, Python libraries such as Requests, Beautiful Soup are utilized (Mallick *et al.*, 2018). After obtaining data, data is preprocessed utilizing Python programming language. This includes tokenizing, excluding unicode errors, stop-word removal for relevant tasks. For implement the classification, PyTorch package, and Hugging Face Transformers library are utilized.

To perform clustering after classification, the Python sklearn package is utilized. Thereafter, the research focuses on generating summaries for the identified clusters. For that, open-source

pre-trained models available in open source LLMs directly Ollama such as Llama3, Mistral AI, and DeepSeek R1 are applied. Ollama enables users to download and run open-source LLMs for free on their computers. At the end, the output of the research is expected to be presented in the form of graphical user interface (GUI). It is created utilizing the Python library, Streamlit. Streamlit is an open-source Python framework for creating interactive data apps.

1.8 Structure of the Dissertation

In chapter one, the project overview, motivation, research problems, objectives, questions, background of the study, and significance of the research are introduced. Apart from those, scope of the study, and feasibility study are also introduced. The aim of this chapter is to give a sufficient number of details about the context and the background of the research study. Further, highlighting the researcher's motive behind selecting this research topic is also done in this chapter.

In chapter two, literature related to the research topic is discussed. Accordingly, literature related to document classification, document clustering, text summarizations, and techniques to improve text summarization are discussed. The aim of the chapter is to explore the current attempts of research and identify existing research gaps that are to be addressed via this study. Further, identifying pros, and cons of current studies to evaluate them, and giving researcher's own thoughts on those studies are also done in the literature review chapter.

In chapter three, proposed methodology is presented with giving reference to methodologies to previous studies conducted related to the research topic. The aim of this chapter is to present the mechanism utilized in the study to find solutions to the research problem, and questions. In other words, the chapter focuses on establishing the methodology to achieve set research objectives. Further, justification is also provided for applying particular techniques in this chapter.

In chapter four, the evaluation and results of the study are discussed. After applying the proposed methodology on the dataset, the obtained results are evaluated and discussed in this chapter. Hence, the aim of this chapter is to evaluate and discuss the results of the research study related to each task. Thus, evaluation of classification of online news, grouping of online news, and summaries generated are discussed in this chapter by utilizing tables, graphs, and other illustrations.

In chapter five, conclusion and future work are discussed. The aim of this chapter is to do a higher-level discussion on whether the set research objectives are achieved and verify the set research questions are answered. Hence, this chapter plays the role in evaluating the overall study in terms of set research objectives, and questions. Further, this chapter also intends to suggest future research work related to this study. That helps improve the future research studies, as well as identifying potential research gaps in future too.

Finally, the appendices and references related to the research are presented. The aim of that is to list down all the references, so that readers can access and obtain detailed information about the previous research conducted related to the research topic. Further, it provides evidence about existing literature. Appendices section is included with code snippets mainly for document classification, document clustering, document summarization. Further, summaries generated by LLMs are also presented in appendices. Additionally, the images of the GUI are also presented in this section.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

The objective of this chapter is to identify, review, and synthesize existing literature related to the selected research area. It is important to identify existing research efforts related to the topic as it gives the opportunity to explore current research findings, identifying potential areas to conduct further research, and making suggestions for future research. Further, reviewing previous literature allows define clear research scope for the current study.

In the body of the chapter, literature is presented in thematic basis as the research involves combining different techniques related to text analytics such as document classification, document clustering, and abstractive text summarization. The chapter begins by exploring literature related document classification which provides the foundation to filter set of documents before clustering them. Then, the discussion moves on exploring literature related to document clustering which provides the basis for grouping news before generating abstractive text summary for each group. Thereafter, the chapter seeks to explore literature related to abstractive text summarization as the outcome of the research is to generate abstractive text summary for online business news. Finally, the research gap is presented which is the basis for progressing with current study.

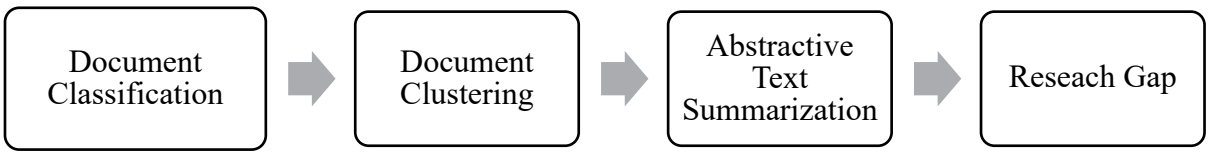


Figure 2.1 Literature review structure

2.2 Document Classification

Document classification is a method of assigning text documents into pre-defined set of classes. Since this is a classification method, it requires a set of labeled text documents with pre-defined

classes to train a document classifier. In document classification, a set of given documents are categorized into pre-defined classes based on the content available in documents. Hence, this is a useful method of information retrieval. Further, this approach enables selective identification of relevant documents from vast number of documents, forming the basis for selecting suitable documents for a specific purpose. There are two categories of document classification: conventional document classification methods, and neural network-based classification methods.

Ahmed and Ahmed (2021) employed conventional document classification methods for classifying of online news articles. The research utilized 75,000 online news articles from Huff Post spanning from 2012 to 2018, covering eight categories: entertainment, crime, politics, business, world news, sports, media, and technology. Despite the considerably larger dataset, reliance on a single news source may constrain the model's ability to learn diverse writing styles, potentially impacting its generalizability. The dataset was experimented utilizing several classification methods, including Naïve Bayes (NB), logistic regression (LR), support vector machine (SVM), and k-nearest neighbors (KNN). The dataset of the study was split into 70% for training and rest for testing. The results of the study suggested that NB classifier recorded highest accuracy of 93%, making it the best classifier for the given task. Further, this was confirmed by other evaluations such as precision (0.93), recall (0.93), and F1 score (0.9). Given the imbalanced nature of the dataset, it is important to employ other evaluation metrics to ensure the generalizability of the results. Notably, LR and SVM had also recorded higher accuracy along with other performance measures mentioned above, making both models good candidates for news classification. However, KNN classifier recorded poor results making it less suitable for news classification task. Additionally, researchers noted that NB achieved the highest performance on imbalanced datasets. This makes NB a strong candidate for practical applications, as real-world datasets are often imbalanced. Nevertheless, employing conventional methods for categorizing news requires rigorous tasks such as feature engineering. The paper has not mentioned sufficient details about the feature extraction such term frequency-inverse document frequency (TF-IDF). Hence, conventional methods require appropriate feature selection and engineering to obtain higher performance of the model. This is one of the reasons neural network-based classification methods such as BERT becomes more popular as they do not require feature engineering.

Daud *et al*. (2023) utilized conventional machine learning (ML) techniques to categorize online news articles from Reuters, containing 40,063 English news articles. The news categorization was for four categories: domestic news, political news, top news, and world news covering

news from 2016 to 2018 years. Most importantly, the research ensured balanced dataset with 4,180 articles for each category by selecting news from the initial Reuters news dataset. This made the final dataset with 16,720 news articles with discarding the rest from the initial 40,063 news articles. This is an improved approach to overcome the dataset imbalance observed in Ahmed and Ahmed (2021). However, the reliance on single data source may limit the model's ability to learn different writing styles. Hence, combining another news source is a better approach to make the results more generalizable. The study utilized SVM along with five other conventional machine learning techniques: stochastic gradient descent (SGD), random forest (RF), LR, KNN, and NB which is mostly similar to the approach of Ahmed and Ahmed (2021). During dataset preparation, the headline and body of the news articles were concatenated and stored together. This is an important task as the news headline plays a significant role in conveying the essence of the news content. The researchers utilized TF-IDF method to extract features prior to applying the algorithm. The results of the study suggested that hyperparameter-optimized SVM algorithm performed well for news classification as opposed to the results of Ahmed and Ahmed (2021) where NB was suggested better over SVM. This could be due to non-utilization of hyperparameter tuning for SVM by Ahmed and Ahmed (2021). Evidently, the findings of the study also proposed that SVM without hyperparameter tuning performed worst for the classification. However, the conclusion has been drawn based purely on the accuracy of the model due to utilization of balanced dataset. Nevertheless, it is still essential to utilize other metrics such as precision, recall, and F1 score, as done by Ahmed and Ahmed (2021), to obtain more generalizable results and compare the results with other studies. Additionally, the research suggested that linear kernel for SVM was more appropriate. The resulted model has also been tested on a new dataset ag_news and SGD performed slightly better than SVM. This could be due to the lack of diversity in the initial dataset, which lead to higher performance with SVM on that specific dataset but lower performance on a different dataset. This underscores the importance of incorporating multiple sources to enhance the generalizability of the findings. Despite the higher performance of SVM, the need for feature engineering and domain knowledge makes conventional models less attractive for news classification compared to neural network-based classification methods such as BERT.

Nugroho, Sukmadewa and Yudistira (2021) conducted a study to explore the efficacy of large-scale news classification by utilizing BERT model. The dataset utilized in this study comprised of news articles from AG dataset which is a commonly available English news article dataset. For the purpose of research, dataset of 120,000 training samples and 7,600 test samples were employed from AG dataset. The utilization of considerably large amount of dataset provides

robust basis for the classification task. The methodology of the research employed five pre-trained models with a different number of parameters, including BERT-Tiny, BERT-Mini, BERT-Small, BERT-Medium, and BERT-Base. Utilizing transformer-based model for classification is an appropriate choice as models such as long-short-term memory (LSTM) does not accommodate parallel processing of text data. Further, it eliminates the need for feature engineering due to its neural network-based architecture and does not require training from scratch, as it is a pre-trained model. The researchers have integrated Spark NLP, an open-source text processing library for training the model. Accordingly, the results of the classification have been compared between BERT models trained with and without Spark NLP. Despite the higher computational overhead, BERT, without the use of Spark NLP, achieved a higher accuracy, with an average accuracy rate of 0.9187. This figure represents the averaged accuracy across the above mentioned five BERT models. Conversely, BERT models utilizing Spark NLP pipelines demonstrated a slightly lower average accuracy of 0.8665. Among the BERT models, highest accuracy was observed for BERT-Base model, both with and without Spark NLP, while BERT-Tiny model recorded lowest accuracy. Overall, the research underscored the potential of BERT-Base model for large scale text classification tasks while advocating for the adoption of transformer-based architectures.

Song (2022) explored the ability of utilizing pre-trained language models on text classification tasks for determining whether the content of a given article is suitable for brand advertising. The research employed a Norwegian-language news dataset containing 3,600 news items covering a period from February 1st, 2017 to January 31st, 2022. Further, each news was labeled for eleven categories such as arms and ammunition, online piracy, terrorism which are considered sensitive to brand safety. Furthermore, there was an exclusion category where a particular news does not fall under any of the eleven categories. The study employed five-pretrained BERT models, including one multilingual BERT and four language models pre-trained specifically on Norwegian. Those four models were BERT multilingual base model (cased), NB-BERT-base, NB-BERT-large, NorBERT and NorBERT2. The dataset of the study was split into 80% for training and rest for testing purposes. The training was done for 20 epochs with base learning rate 2e-5. Furthermore, it is important to highlight that the researcher employed the truncation method since BERT can process only 512 tokens at a time. However, the researchers suggested that the initial part of the news articles often contain the most information. Hence, input the initial news content is an optimal choice for BERT model. The results of the study suggested that NB-BERT-large achieved the highest performance, with a receiver operating characteristic curve (AUC-ROC) value of 0.9492. Additionally, research

suggests that higher number of parameters ensure higher performance of the model. Accordingly, the research advocates the utilization of transformer-based models for news classification over traditional methods due to limitations of traditional models such as handcraft feature engineering. However, since the news dataset was in Norwegian, hence, there could be limitations in generalizing the findings to English news datasets.

The study of Padalko *et al*. (2023) focused on exploring the application of the BERT model for detecting fake news. In other words, the aim of the research was to utilize the BERT model to differentiate between factual reporting and misinformation. The dataset utilized in the study was "Fake-Real News", an English news dataset from Kaggle. Accordingly, the study employed a binary classification to distinguish between 'Fake' and 'True' news. The results of the study achieved accuracy of 79.88% and an AUC-ROC value of 0.87. However, the research revealed fluctuations in testing accuracy and a tendency to misclassify true news as fake, highlighting areas for future improvement. Hence, it is important to ensure consistent testing accuracy when utilizing BERT model for news classification. However, testing loss is also another important fact to observe overfitting of the model. The study has not revealed information about that. Further, the results revealed better performance of the model in identifying the real news class than the fake news class. Further, study illustrated model's decline in training loss across 15 epochs as additional training yielded minimal to no improvement in reducing loss. This suggests that 15 epochs can serve as an upper bound for training a model, in contrast to the 20 epochs utilized by Song (2022). Additionally, study highlights the importance of utilizing a balanced dataset to ensure accurate classifications.

Chen (2024) performed a news classification study on news headlines based on the BERT model. The research employed 2,000 news headlines in English news from sources such as Xinhua, China Daily, China News, Oriental, Sohu News etc. The utilization of diverse sources is an important aspect as the model can learn different writing style. In other words, diverse sources allow expose the BERT model for diverse examples which in turn increase the model's adaptation for diverse examples. On the other hand, Song (2022) suggested that initial part of the news contains most important information. Hence, utilizing the headline of the news, as it represents the initial part of a news article, is justified. However, the headline alone may not be sufficient, and it may need to be combined with the first few sentences of the news to enrich the content for the classification task as done by Daud *et al*. (2023). The study involved multi-class classification, as it targeted to categorize news headlines into four categories: social, political, economic, and scientific/technological news. Notably, the dataset of the study was split into 70% for training and rest for testing. The experiment was conducted between 10-20

epochs. The findings of the study demonstrated that an accuracy of 0.7 was achieved after 10 epochs with a learning rate of 5e-6, while 20 epochs with a learning rate of 1e-5 resulted in an accuracy of 0.6. This suggests 10 epochs can serve as the minimum for training the model which is lower than 15 epochs (Padalko *et al*., 2023), and 20 epochs (Song, 2022). Apart from the accuracy, the results of the study have been supported by metrics such as training loss value, and test accuracy and test loss value. Particularly, declining value of testing loss suggests that the model is not overfitting. Further, the researcher suggested that adjusting the learning rate parameter and ensuring a balanced dataset are important for enhancing the model's generalization ability, which is similar to suggestions made by Padalko *et al*. (2023).

2.3 Document Clustering

Document clustering is a method of grouping corpus of documents based on features by utilizing an unsupervised machine learning algorithm. Generally, it results in documents that are similar and related to the same group or cluster. Different methods have been adopted to perform document clustering such as hierarchical clustering models, centroid-based clustering models, distribution-based clustering models, and density-based clustering models. Grouping news documents helps gain more meaningful insights while uncovering hidden patterns in news data.

Kumbhar *et al*. (2020) explored the ability of utilizing dimension reduction techniques to reduced number of features to enhance the performance of document clustering techniques. Accordingly, the research utilized two dimensionality reduction techniques, singular value decomposition (SVD) and non-negative matrix factorization (NMF), each combined separately with K-means clustering to categorize the news. The dataset utilized in the study was 20 newsgroup, an English dataset from Kaggle that consists of 1,000 text documents, which are divided into six sections: computer and electronics, sports, science and medicine, religion, politics, and automobile and e-commerce. The research employed preprocessing of stop word removal which ensure removing of words with minimal value. This is an importance aspect when processing large number of text documents as it helps reduce computational cost. The research employed TF-IDF to extract features from the documents. Document clustering was performed as K-means, SVD K-means, and NMF K-means across different values of K. However, study of Marutho *et al*. (2018) utilized elbow method to decide the optimal K-value for clustering news headlines. Hence, utilizing a method such as the elbow method to determine the optimal K-value, along with testing different K values, could have given better insights into the results. The results suggested that techniques with dimensionality reduction outperform the

usage of K-means alone, based on homogeneity score, completeness score, and adjusted rand score. However, those metric values were considerably low even for a dataset of 1,000 text documents, indicating possibilities for further improvement. Additionally, utilizing a metric such as silhouette score could help gain more insights on the quality of the clusters generated. The researchers suggested application areas for K-means clustering and finding similar documents was one of them. This is an important suggestion, as grouping similar documents supports text analytics tasks such as abstractive text summarization. By grouping similar news together, it becomes convenient to summarize utilizing modern methods such as large language models.

Alternatively, agglomerative hierarchical clustering, a hierarchical clustering model, can also be employed to cluster documents. Adek, Dinata and Ditha (2021) utilized that method to cluster Indonesian language online news in Aceh, a province in Indonesia. The purpose of this study was to cluster the documents and identify keywords of each cluster. Then, whenever a user inputs a keyword, display the news for the clusters where the keyword belongs to. To identify the keywords, the researched utilized a training dataset of 90 documents, representing different news categories, including economic, sport, political etc. For the purpose of testing, 1,000 online news documents from 10 online news websites in Aceh from July 2016 to March 2017 were employed in this study. Similar to Kumbhar *et al*. (2020), TF-IDF was applied to extract features from the documents. The performance of clustering was measured by accuracy where the algorithm groups the news correctly into correct cluster based on keywords. The results suggested higher grouping performance of news with average accuracy of 89.84%. Nevertheless, agglomerative hierarchical clustering method is computationally expensive which is a main disadvantage when dealing with large datasets. Therefore, it suggests a valid reason for utilizing K-means method. Additionally, the clustering method was evaluated utilizing the accuracy of the output generated for user input query. Instead of relying solely on accuracy as the measure, it is preferable to measure the quality of clusters alongside a more appropriate metric such as the silhouette score. Additionally, implementing a different feature extraction method such as word embeddings and comparing the clustering results is an improved way to conduct this research.

Shevendrakumar (2023) conducted research to perform document clustering utilizing several methods, namely agglomerative clustering, K-means, and DBSCAN. Adopting three different methods, including the methods applied in previous studies of Kumbhar *et al*. (2020), and Adek, Dinata and Ditha (2021) to cluster news for same dataset is an ideal approach to find a better method. The study utilized 21,578 documents from Reuters news covering various categories

such as business, politics, and sports. Utilizing real-world news data in this study, with a larger sample size is a key highlight of this study. Accordingly, the dataset employed in this study was considerably larger than the dataset utilized in previous studies of Kumbhar *et al.* (2020), and Adek, Dinata and Ditha (2021). Further, it is important to note that this study considered metadata of news documents such as author, date, and topic code which is important to get more idea when describing clusters. The feature generation was done utilizing TF-IDF method which was similar to previous studies Kumbhar *et al.* (2020), and Adek, Dinata and Ditha (2021). Results of each clustering method was evaluated utilizing "Adjusted Rand Index" (ARI) where ARI closer 1 indicates better clustering. The results of the study suggested that agglomerative clustering method yielded higher ARI, followed by K-means, and DBSCAN. However, incorporating additional measures utilized by Kumbhar *et al*. (2020), such as the homogeneity score and completeness score, could provide better clarification about cluster quality. Additionally, incorporating the silhouette score could provide more clarity on cluster quality. Nevertheless, researcher still suggested that all three algorithms reflected average performance. Despite the better results, agglomerative clustering is computationally expensive method and does not provide enough scalability. Therefore, there is still potential for utilizing K-means clustering in news clustering, given its scalability, fast and easy to understand nature as suggested by Bano *et al.* (2018). Further, the researcher suggested employing domain-specific knowledge to enhance feature extraction, and to utilize more diverse dataset. Furthermore, a diverse dataset could contain more valuable information, which is suitable in the domain of business to gain valuable insights. Hence, utilizing a dataset containing news from multiple sources is an appropriate way to diversify a dataset. Moreover, employing an alternative feature extraction method such as word embeddings and comparing the clustering results is a more effective approach to conduct the research.

2.4 Abstractive Text Summarization

Automatic document summarization (ATS) or often referred to as text summarization, is the processing of generating condense version of text from large amount of text applying computational methods. Hence, a summary is a short, accurate, and fluent form derived from large amount of textual data. In general, ATS techniques are divided into two categories, namely extractive techniques, and abstractive techniques. Extractive methods aim to generate a summary by extracting some key subset of content from a document that contain core information (Sarkar, 2019). Hence, the summary includes extracted sentences from the original document. On the other hand, abstractive methods aim to summarize document by leveraging natural language generation techniques (Sarkar, 2019). Hence, the summaries generated by

abstractive methods include new sentences which is not there in the original document. The ability of abstractive methods to generate new sentences makes it more suitable for summarizing large number of documents over extractive methods. This is because applying extractive methods on large number of documents create higher number of extracted sentences in the summary which in turn makes it difficult to comprehend by human.

Tarannum, S. *et al*. (2021) conducted research to reveal approaches, and challenges to NLP based text summarization techniques. The research attempted to conduct a systematic literature review on current methods applied to text summarization based on research articles from 2015 to 2021. Related to abstractive text summarization, the researchers suggested "Long Short-Term Memory" (LSTM) architecture as the suitable method for encoding and decoding data. Thus, the study suggested that bi-directional LSTM layer is more suitable method for encoding data from the source document, and one-word-at-a-time LSTM layer as the appropriate method for decoding data or creating summaries. However, it is important to note that one of the drawbacks of the LSTM architecture is its inability to process sequential text data parallelly. To solve that, the study proposed utilizing attention mechanism to improve the LSTM architecture. Despite that, currently, the focus moves towards applying transformer-based architecture for natural language generation (NLG) tasks such as abstractive text summarization by addressing the drawbacks of LSTM architecture. Nevertheless, similar to the findings of Moratanch and Chitrakala (2016), this study also proposed that the abstractive methods accommodate advanced capabilities such as paraphrase, generalization, and integration of real-world knowledge, despite the convenience of extractive methods. Most importantly, the study highlighted that deep learning has made abstractive methods more successful, providing clear evidence for utilizing those methods for summary generation such as transformer-based techniques. However, the study does not report about change of focus on applying transformer-based architecture for abstractive summary generation.

Wang *et al*. (2021) suggested an unsupervised graph-clustering framework called "FinGC" to perform financial news summarization with special focus on stock market. The study utilized pre-trained models BERT, and PEGASUS for sentence embedding and all the sentences from financial news were embedded by utilizing those. Application of BERT for sentence embedding is suitable as BERT is considered a better method for natural language understanding (NLU) or encoding text data. Then the study utilized graph embedding method to represent the financial news, and clustering was performed by applying K-means clustering algorithm. The choice of K is based on heuristic search method. It is justifiable that utilizing different K values and trailing the number of clusters. However, applying elbow method to decide the K value is

also a suitable method (Marutho *et al*., 2018), which this study has not considered. The study utilized 72,476 news instances from eastmoney.com for the year 2017. For the comparison purpose, the researchers also employed benchmarking extractive summarization methods such as Lead-3, LexRank, RASR, TCSum, HNet, and abstractive summary methods such as flat transformer (FT), hierarchical transformer (HT), FT_{bert}, HT_{bert}, GraphSum. This is important as it allows comparison between abstractive and extractive methods. Based on the recall-oriented understudy for gisting evaluation (ROUGE) metrics such as ROUGE-1, ROUGE-2, the suggested FinGC model outperformed other extractive, and abstractive methods. However, to calculate above metrics, reference summaries generated by human are required, making it less suitable when such summaries are not available. Further, this was further confirmed by human evaluation of the generated summaries too. Overall results of the study suggested that employing pre-trained language models improve the generated summary. Utilizing human evaluation is important, however, when the amount of textual data getting larger, the ability of human to comprehend data and generate summaries become limited. Hence, incorporating human evaluation may not be as suitable with larger number of textual documents. Further, the model predominantly focused on supporting investors. Hence, it limits the usability of the suggested model in different use cases such as company decision making related to launching new products, business trends etc. Nevertheless, the study was focused more on NLU or encoder side rather than on NLG or decoder side. The advancement of new architectures based on transformers has proven better results in NLG or text generation tasks. Therefore, it is important to explore summary generation by applying those methods and compare the results.

The suitability of applying deep learning methods for abstractive text summarization has motivated the researchers to explore that by employing transformer neural networks. Patel (2022) conducted a study on financial news summarization by applying transformer neural network. The research was conducted utilizing 2,224 news articles from BBC covering sectors such as business, politics, tech, and entertainment. For text representation, the researcher employed Word2Vec algorithm or word embeddings which is more suitable to capture semantic and contextual meaning of language. The study employed GPT-2 model/transformers by Hugging Face, an open-source pre-train model to generate summaries. The choice of GPT model as the decoder or NLG ensures improved approach for summary generation in comparison with the study of Wang *et al.* (2021) where the focus was on NLU than NLG. In conclusion, the researcher suggested that the transformer-based model generated better summaries than Seq2Seq LSTM model. However, that conclusion was drawn based on the evaluation metric called “loss function” which measured how the summary improved over

epochs. This is in contrast with the evaluation metric ROUGE toolkit, suggested by study previous literature (Moratanch and Chitrakala, 2016). However, to utilize the ROUGE toolkit, human generated reference summaries are required. Nevertheless, as mentioned by Moratanch and Chitrakala (2016), utilizing generalized framework is a key challenge to utilize abstractive methods.

As suggested by the previous study of Tarannum, S. *et al.* (2021), Li *et al.* (2023) explored the feasibility of employing bi-directional LSTM encoder, and attention-based decoder to summarize financial news. The study proposed an enhanced Seq2Seq model named TLGA for financial news summarization with special focus on predicting stock price movement. The model was tested with financial news in Chinese and English. Further, the model also utilized attention-based mechanism that was suggested by Tarannum, S. *et al.* (2021) to improve the performance of encoder and decoder when utilizing LSTM architecture. The model considered impact of past news on current news and latent casual relationships among news. Related to news, latent casual relationships are the cause-and-effect connections between news. Financial news contains latent casual relationships, and hence, considering past news is important to predict the stock prices. LSTM is capable of capturing long terms dependencies of data. Hence, it justifies the utilization of LSTM encoder in the study. Further, the study emphasized lack of large-scale financial corpus. As a solution, the researchers created a new large-scale Chinese financial news summarization dataset (LCFNS), containing 430,820 samples from several major financial portals including East Money, Sina Finance, China Business News from January 2013 to June 2020. Most of the previous studies relied on reference summaries generated by others. The creation of a summarization dataset specific to the task was not explored in those studies. Hence, this approach in the study ensures better evaluation of summaries, as earlier studies such as Patel (2022) were limited in their ability to apply metrics such as the ROUGE toolkit. Apart from that, two other benchmarking datasets, namely LCSTS dataset, and CNN/Daily Mail dataset were also utilized. Similar to study of Wang *et al*. (2021), the researchers utilized ROUGE toolkit to evaluate the quality of the summaries, despite the drawbacks of ROUGE to evaluate summaries suggested by Zhang *et al.* (2024) such as ROUGE's tendency of penalizing summaries that utilize different wording from the reference summaries. Notably, BARTScore was also utilized to assess informativeness and faithfulness of summaries which was an enhanced approach compared to purely relying on ROUGE toolkit. The proposed TLGA model outperformed all other methods based on BARTScore. Further, TLGA model generated better results than all other models based on ROGUE 1, and ROUGE-2 for LCFNS dataset, and Fin-LCSTS dataset (this a dataset containing 209,092 examples

related to financial domain created from original LCSTS). However, BART model performed better than proposed model based on ROUGE-L, a metric focus on evaluating the readability of the summary. This implies that utilizing open-source models such as BART is still suitable for text summarization tasks. In contrast, for benchmarking dataset CNN/Daily Mail, models such as UniLM, and BART performed better than proposed TLGA model in terms of ROGUE-1, and ROUGE-2, ROUGE-L. Nevertheless, the researchers also utilized human evaluation of summaries and suggested that the proposed TLGA model was better. Despite the positive confirmation from human evaluation, and better performance based on ROUGE in some cases, utilizing open-source models such as BART has still generated better results. This indicates researchers still can rely on open-source models. The advancement of technology has led to the development of many free open-source models, such as those available in Ollama, a platform offering language models that can be downloaded and run on personal computers. Further, the focus of the research was on stock market. However, the focus can be broadened by increasing the usability for any parties who is interested in the field of business such as a management of the company, board of directors etc. Moreover, the summary was generated for individual news articles which is a drawback when a user wants to comprehend news from large number of news articles. This is because summary for each news for large amount of news articles could overload the user. Additionally, utilizing transformer-based models gained considerable attention for text summarization recently. Thus, utilizing transformer-based decoder could enhance the performance of the suggested model by this study.

The higher performance of transformer-based model such as GPT based models to generate summaries lead to explore performing that task by utilizing LLMs. Manathunga and Illangasekara (2023) proposed a retrieval-augmented-generation-assisted representative vector summarization (RVS) approach called “docGPT” related to the field of medical education. The “docGPT” is a document intelligence program written in Python utilizing LangChain framework. The study attempted to mitigate the risk of hallucination by utilizing RAG technique. The study utilized various textual sources related to medical education such as PDFs, text documents, spreadsheets and slide presentations ensuring diversity of the dataset. The textual data from images and scanned PDF files were extracted utilizing optical character recognition (OCR). Before generating summaries, the documents were clustered into groups utilizing K-means clustering method. This ensures the summary generated captures the diverse aspects of the original document/s well. However, the research did not employ metrics such as ROUGE score to evaluate the summaries as it may not have human generated reference summaries. Instead, evaluate the summaries by comparing responses to queries with RAG and

without RAG. In conclusion, study suggested that proposed docGPT generates more accurate answers whereas ChatGPT (without RAG) generates more generic answers. The researchers highlighted the importance of applying RAG technique to improve the quality of summaries and minimizing hallucinations in other domains too. This suggests the potential applicability of LLMs for generating summaries along with RAG technique in the business field for various purposes. However, utilizing ChatGPT presents few limitations such as including token constraints and increased costs associated with processing large volumes of text data which makes it less attractive for experimental nature research. Hence, free open-source models, such as those available from Ollama, could serve as a better alternative, as they are free, despite their limited capabilities when utilized with low-processing power computers.

Tripathi and Pandey (2024) employed Hugging Face transformers library to generate summaries for BBC news articles covering five categories business, entertainment, politics, sport, and tech. Employing transformer-based architecture is in line with suggestions made by previous study of Patel (2022) to improve the quality of the summary. However, the number of news articles or the sample size is not mentioned in the research paper. Nevertheless, the researcher utilized configuring parameters such as max length, minimum length which makes the summary generation more versatile and flexible by changing parameter values. Moreover, the research utilized different categories of news that makes the dataset more diverse. The parameter adjustment is useful to generate summaries based on the category of the news as some categories require shorter summaries whereas other require longer. As suggested by prior study of Moratanch and Chitrakala (2016), the researcher utilized ROUGE metric to evaluate the quality of summaries. However, the numerical figures for ROUGE were not available in published research paper. Nevertheless, the researcher concluded that utilizing Hugging Face transformers library generated summaries efficiently by capturing essential information while upholding coherence and relevance. However, there could be limitations of Hugging Face transformers library with large textual data, as it may fail to process the data after reaching a certain threshold, despite being free. This makes this approach less attractive for experimental research compared to free open-source models available in Ollama. Further, the application of categorizing or clustering news, and summarizing news for those categories or clusters would be an improvement that can be made to improve this study.

Amidst the rising awareness of abstractive summary generation by employing LLMs, Keswani *et al*. (2024) proposed an efficient approach to retain context over extensive texts or multiple documents with special reference to summarization and question answering. The study utilized Meta's Llama2 model which has a maximum token limit of 4096. Despite the effective

summary generation, token limitation is a drawback of applying LLMs for summary generation. Similar to previous study of Manathunga and Illangasekara (2023), the approach of the study clustered the documents using K-means before generating summaries. The number of clusters or K was decided utilizing elbow method. Further, the study limited the cluster count between 3-15, given the limitation of LLM providing summary beyond the threshold. The summary generated for each cluster was combined at the end to generate final summary. However, in case of a domain such as business, it is important to know about each cluster rather than combined summary for all. Hence, generation of summary for each cluster is a potential area to be addressed in future research. Despite the advantages of employing LLMs for summary generation, hallucination is a possible issue of LLMs. However, the researchers applied RAG to handle that issue by feeding more contextual information for summary generation process which is similar approach to Manathunga and Illangasekara (2023). Finally, RAG was evaluated by utilizing RAGAS tool that considers context relevancy, faithfulness, answer relevancy, and context recall. The results recorded a context relevancy score of 0.42, a faithfulness value of 0.99, an answer relevancy value of 0.93, and a context recall value of 0.55. RAGAS is a tool that allows for evaluating summaries in some aspects without the need for human-generated reference summaries. However, when applying it to large volume textual data, the limited processing power of computer could restrict its capabilities. On the other hand, the generated summaries were assessed with reference summaries utilizing the ROGUE score, in line with previous studies of Li *et al*. (2023). In conclusion, research proposed that applying LLMs to generate summaries provides improved results. However, the suitability of the prompts written is also an important aspect of generating summaries through LLMs. On the other hand, the ability to experiment different prompts makes the LLM-based techniques more versatile, allowing users to employ for customized purposes.

Similar to the research of Manathunga and Illangasekara (2023), Alkhalaf *et al.* (2024) explored the possibility of utilizing LLMs to extract key clinical information from a large volume of data in electronic health records (EHR). The study consists of two main tasks: utilizing LLMs for summarizing nutrition care, and extracting malnutrition risk factors from the provided data. The data includes both structured and unstructured EHRs related to malnutrition management across 40 Australian aged care facilities (RACFs). The popularity of LLMs is growing for NLP tasks such as information extraction and abstractive text summarization. Hence, this research presents an appropriate use case for applying LLMs for text summarization in the medical field similar to Manathunga and Illangasekara (2023). However, the research employed the Llama 2 model from Huggingface library, in contrast to Manathunga and Illangasekara (2023), who

utilized GPT-3.5-turbo, and but similar to Keswani *et al*. (2024). Nevertheless, both GPT and Llama 2 are transformer-based architectures that have demonstrated improved performance in text generation over other architectures such as LSTM with the ability of parallel processing. The study also utilized zero-shot prompt engineering where an AI model is given with a task or question without any prior examples or specialized training for that particular task. This is a more appropriate technique for domain specific tasks. The important fact about this study is the application of RAG as a method of improving the information extraction and summarization as well as mitigating hallucination. The results of the study suggested that generative AI model is effective in summarizing and extracting information about nutritional status of RACFs' clients. Further, summaries provided concise and accurate representation of the original data with an overall accuracy of 93.25%. The accuracy has been measured by comparing model's output with manually labeled dataset (ground truth), which was initially assessed by three nursing experts, achieving 92% agreement. This is an important step in this research as obtaining the support of domain experts to assess the labeling of data. Further, incorporating RAG improved the summarization process, boosting accuracy by 6%, from 93.25% to 99.25%. This highlights the importance of utilizing of RAG to enhance text generation by LLMs. Despite the overall performance, the research has observed cases of hallucination where the model generates inaccurate answers without following the prompt. Further, the study reveals that the model fails to generate consistent output when faced with minor variations in the prompt. Additionally, without purely relying on accuracy, the evaluation of output can further be measured by answering questions from the model and evaluating them based on criteria such as faithfulness, answer relevance, and context relevance proposed by Es *et al.* (2023).

2.5 Research Gap

The previous literature referred above such as Patel (2022) and Tripathi and Pandey (2024) demonstrates that some of the abstractive summary generation tasks has been conducted without combining with other methods such as clustering. Even in studies that combine both techniques such as Keswani *et al.* (2024) and Wang *et al*. (2021), cluster-wise summary generation is not explored to the best of the knowledge. Additionally, in the domain of business, some business news is more important than others. To decide that there is a room for exploring to apply pre-trained model for news classification. To the best of the knowledge, no existing studies have combined news classification with clustering and abstractive text summarization techniques.

On the other hand, previous studies such as Patel (2022) and Shevendrakumar (2023) adopted single source of data such as BBC news, and Reuters news respectively. Further, Shevendrakumar (2023) suggested employing more diverse dataset. Hence, this can be improved by utilizing a dataset created by combining multiple sources. Further, that ensures capturing diverse news and languages styles which is an important fact in the domain of business to gain insights from textual data.

Finally, the emergence of open-source free tools such as Ollama has enhanced the capability for experimentation, providing access to various LLMs such as Llama, Mistral, DeepSeek etc. Hence, there is an opportunity to explore the effectiveness of summary generation by employing those LLMs. Altogether, the combined approach of document classification, document clustering, and abstractive text summarization is the novelty of the proposed study.

CHAPTER 3
METHODOLOGY

3.1 Introduction

The objective of this chapter is to outline the methodology that is utilized in the research study. This is the chapter that highlights the methods that is employed to achieve the set research objectives initially. The aim of the research is to suggest an approach to gain meaningful business insights from online English business news in Sri Lanka. Hence, the methodology includes techniques to select relevant business news, techniques to group the news, and techniques to generate summary for each group to unveil business insights to support business decision making.

The initial focus of this chapter is to provide detailed description about research design in terms of research philosophy, research type, research strategy, time horizon, sampling strategy, and data collection. Then, methods related to data pre-processing, feature extraction, news classification, clustering, and abstractive text summarization are presented. Thereafter, the chapter is also be presented with architectural diagram, and the limitations of the study in terms of sampling, and other resources. Finally, a concluding summary about the chapter is presented.

3.2 Research Design

3.2.1 Research Philosophy

The research philosophy is related to the belief about the way in which data about a phenomenon should be gathered, analyzed, and utilized. The aim of this study is to gain business insights from online English business news in Sri Lanka. To achieve that, statistical, computational techniques and tools are utilized while making interpretations of the results obtained via those techniques and tools. Hence, this is more related to the belief that knowledge is obtained by doing and acting. Further, there are one or more solutions to gain business insights from online business news (ontology). On the other hand, to gain business insights, knowledge should be measured and interpreted (epistemology). Overall, the focus of the research is to solve problem by utilizing practical methods to achieve set research objectives. Thus, it can be concluded that this research falls under pragmatic research philosophy.

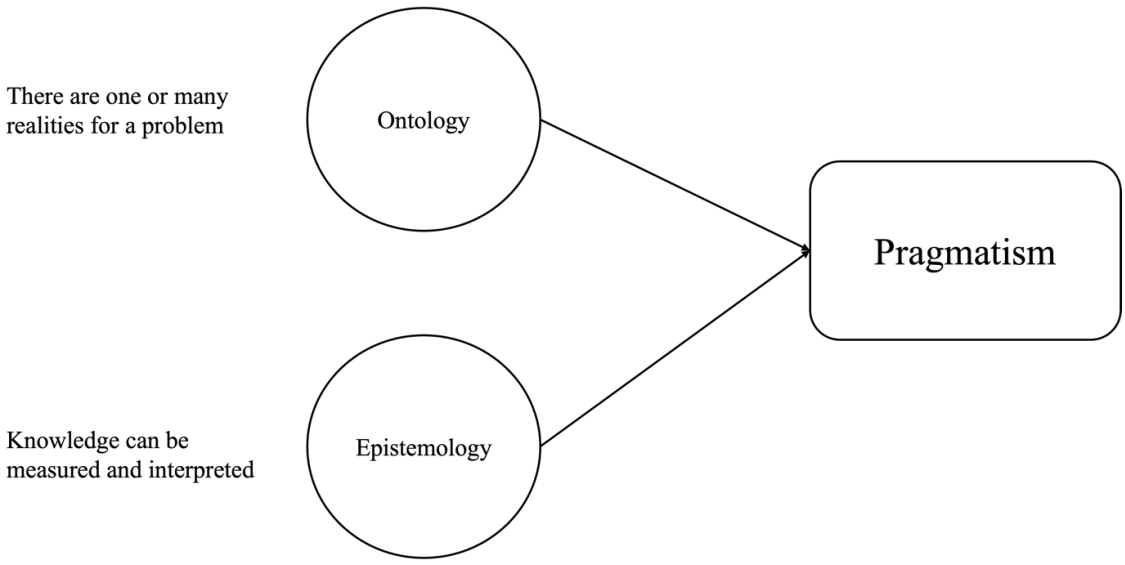


Figure 3.1 Research philosophy

3.2.2 Research Approach

There are two main research approaches namely, deductive, and inductive. In deductive approach, the researcher starts with a theory and develops hypotheses based on that theory, followed by collecting data to test those hypotheses. On the other hand, inductive approach collects data relevant to the topic initially, and analyze them to identify patterns, followed by generalizing the findings by the researcher. Thus, in inductive method, the conclusions of the study are based on specific dataset utilized. This study also starts with collecting data, followed by analyzing them, and generalizing the results based on the dataset. Hence, it can be concluded that this study falls under inductive approach.

3.2.3 Research Strategy

The focus of the study is to examine historical online business news documents to gain business insights. Hence, a most suitable research strategy for the study is archival research strategy where the researcher utilizes a collection historical news documents obtained from online news archives. Further, online archive makes it convenient to access news documents. Particularly, this study's focus on online business news leads to utilize digitized historical news archives. However, it is important to note that utilizing online news may cause a limitation of only capturing fraction of the news articles included in the physical newspapers (Beach and Hanlon, 2022). Online news providers typically keep their own archive for paper news documents and sometimes it

is required to access those data through a subscription service (Beach and Hanlon, 2022). However, some online news archives are open to anyone and allow obtain those data utilizing web-scraping techniques (Beach and Hanlon, 2022). In Sri Lankan context, most of the online newspapers have free access to at least a fraction of the whole newspaper. Hence, those archived news documents in news websites over past years can be utilized as the main sources to collect data for this study.

3.2.4 Time Horizon

The dataset is collected for the study is related to a period of time. In other words, the dataset that is utilized in this study is not collected at specific point of time. Hence, the data can be considered longitudinal data. There are few reasons for collecting data for over a period of time. Firstly, it ensures gathering adequate amount of data for the study. Previous studies of Adek, Dinata and Ditha (2021), Patel (2022), and Shevendrakumar (2023) have utilized dataset with the sizes of 1,000; 2,224; 21,578 news articles respectively. Secondly, to ensure adding diverse data to the dataset as suggested by Shevendrakumar (2023). Thirdly, longitudinal data allows identifying patterns of data over time, which is in line with the current research topic "unveiling patterns of online English business news." Accordingly, the dataset of this study contains online news articles spanning from 2019 to 2024.

3.2.5 Sampling Strategy

The nature of the dataset that is required for this study is online English business news that is available in various sources including online blogs, online newspapers, and websites. Hence, it is required to identify potential data sources that suits the set research objectives. The following criterion are considered to select data sources.

a) **Relevance** – This is to ensure that the data sources align with the research objectives. In this research, a source providing business specific news satisfies the relevance criterion.

b) **Quality** – This is mainly related to selecting sources with credibility, and accuracy. If a particular source is cited, or referred in the field of business, that is an indicator of credibility, and accuracy of a source. Accordingly, utilizing

industry-accepted data sources in the domain under consideration, and utilizing highly cited sources would ensure the quality criterion.

c) **Diversity** – The data sources provide diverse range of language styles, topics, and perspectives. Utilizing only one source might not capture diversity adequately, despite covering news on diverse topics. Hence, utilizing more than one source is appropriate to enhance the diversity of the dataset.

d) **Accessibility** – This is related to data accessibility. The ability to scrape and obtain the data freely can be attributed to accessibility of sources related to this study. It is important to note that some online sources do not allow scraping data, despite being freely available. In such cases, there is a concern of utilizing those sources.

There are many newspapers available in Sri Lanka. Presumably, all of them should have online news portal too. Thus, the more appropriate method is to gather news from all those websites. However, that process is time consuming. Hence, proper sampling method should be utilized to address that to select more appropriate subset of news sources that is in line with set research objectives. The goal of research leans towards unveiling patterns of online business news. Therefore, selecting business news specific newspapers is more appropriate. In other words, selecting newspapers that is more align with research purpose or utilizing purposive sampling for selecting news sources is more suitable. The reason for purposive sampling is to support the research objectives with better matching sample while ensuring rigor and reliability of the results (Campbell *et al*., 2020). Given the research objectives, data sources that are widely utilized and accepted in the domain of business field should be considered.

Stock brokerage firms play a vital role in economy by facilitating investments. Thus, the news sources utilized by them can be considered industry-accepted data sources. Hence, priority is given to online news sources in Sri Lanka such as Daily FT, EconomyNext, and Lanka Business Online that are utilized by stock brokerage firms. This is confirmed by referring sources attached to the reports send by stock brokerage firms to their clients, and communicating with analysts who has worked for stock brokerage, and investment advisory firms in Sri Lanka. Further, these three sources are suggested in top of Google search for “Sri Lanka’s top business newspaper” (Figure

3.2). Utilizing multiple sources ensures the diversity of the dataset. This is in accordance with the suggestion made by Shevendrakumar (2023) to employ a diverse dataset.

On the other hand, company websites also provide news related to respective companies making it more appropriate for this research study. However, utilizing company website as the news source is limited by three facts.

a). Firstly, the frequent availability of news. Company websites for 283 companies listed in Colombo Stock Exchange were checked and based on that, it was evident that overall frequency was low. The company website news releases were least frequent for around 67% companies (191 of the 283 companies), and rest was more frequent. Even among the most frequent companies, banking and finance related companies were predominant. It is understandable that banking and finance sector play an important role in the economy. However, only referring those companies does not give an overall picture about the economy and business environment in Sri Lanka.

b). Secondly, the news in company websites may have included bias statements about respective company as those are written by the company itself. Those bias could be learned by machine learning algorithms, resulting suboptimal outcomes.

c). Thirdly, some company websites had restricted scraping data by utilizing Python libraries Requests and BeautifulSoup. This may be due to concerns such as server overload, copyright infringement etc.

Therefore, by considering all above reasons, researcher decided not to utilize company website news as a source of the news data.

Accordingly, the remaining three sources – Daily FT, EconomyNext, and Lanka Business Online – were assessed utilizing the above-mentioned four criteria. News scraping from EconomyNext was unsuccessful when utilizing the Requests and BeautifulSoup libraries, leading to its exclusion from the list of potential sources. Eventually, Daily FT and Lanka Business Online were chosen as they met all the criteria outlined above, and it is demonstrated in Table 3.1.

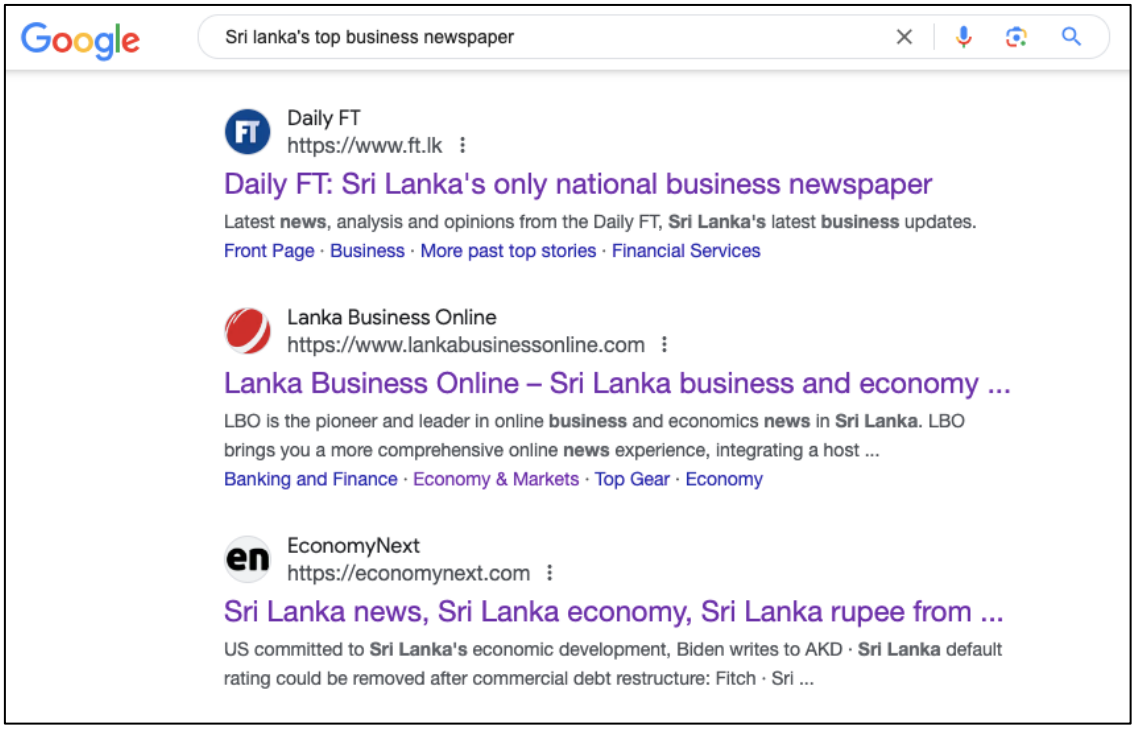


Figure 3.2 Sri Lanka's top business newspaper

Table 3.1 Summary of Assessment of Potential Data Sources

Source	Business specific	Can obtain freely	Overall frequency of news	Scrapable via Requests/BeautifulSoup libraries	Utilized in the study
Company websites	Yes	Yes	Low	Yes/ No*	✗
Daily FT	Yes	Yes	High	Yes	✓
EconomyNext	Yes	Yes	High	No	✗
Lanka Business Online	Yes	Yes	High	Yes	✓

*Some company websites permitted scraping, while others restricted it.

3.2.6 Data Collection and Annotation

a). Data Collection

There are two main methods to collect online news data. One method is data collection through APIs, while other is data collection by web scraping. To utilize API, it is required the data source provider to provide an endpoint with and API key and authentication. However, this study employed news from online news providers in Sri Lanka which did not have API facility for best of the knowledge. Hence, web scraping

was the plausible way to collect data for this research project. Further, Python has packages such as “requests” to send HTTP requests, and “BeautifulSoup” to parse HTML and XML documents. However, these packages are comparatively slower compared to packages such as Selenium, and Scrapy, despite being suitable for low-complexity projects and beginners. Since the researcher’s experience is limited for advanced tools, “requests” and “BeautifulSoup” Python packages were utilized for scraping online business news. The scraped data was stored in comma separated value (CSV) format similar to the approach by Ahmed and Ahmed (2021).

The objectives of the research study include classification of initial set of news documents. This is because some news articles are more relevant to gain business insights whereas some are less relevant. For instance, news about company launching a new product is more relevant whereas a company doing a social project is less relevant to gain business insights. Hence, initial classification is required to remove the less relevant news documents. In other words, it is required to train a classification model to filter more relevant news from a given news dataset. For that purpose, a separate set of news documents (referred as mini dataset hereafter) other than main dataset of news documents was required. The reason for utilizing a different dataset to train classification model was that it would be meaningless to employ the same dataset to train the classification model and perform the classification on the same dataset as the classifier has already seen the data. Further, it is waste of resources to label a dataset and perform the classification on the same dataset. Thus, news documents were scraped for training the classification model apart from the dataset that would be utilized for gaining business insights. Accordingly, the dataset composition for train the classification model (mini dataset), and project input (main dataset) are demonstrated in Table 3.2. The collection ensured a balanced dataset by representing same number of news documents from both sources (i.e. DailyFT, and Lanka Business Online) for both classifier, and input.

b). Data Annotation

To perform a classification task, data annotation should be done for the mini dataset as the first step of building a classification model. Labeled data servers as the ground truth in the study. Ambiguity and subjectivity are two of the challenges in annotating data (Chou *et al*., 2024). In this study, all the news could have some impact on business firms as the news were obtained from business specific news websites. Further, determining

most relevant news is a subjective choice and it contains some element of ambiguity too. Therefore, a logical foundation is necessary to annotate the news as "Relevant" and "Not Relevant". The logical basis for that is presented as follows.

The stock market plays a key role in any economy. Consequently, news articles that influence investors' decisions in the stock market hold significant importance. Hence, the influence of such news on the stock market provides a basis for determining the relevance of news to business decisions. Based on the findings of previous studies, which provided evidence to determine the relevance of news, the basis for annotating/labeling news as "Relevant" or "Not Relevant" is justified as follows.

The study of Yu *et al*. (2021) employed expertise of domain specialists related to stock market to assess the relevance of news articles in relation to their impact on the stock market. This indicated the importance of utilizing domain experts to annotate news data. Accordingly, researcher's personal experience in industry research over seven years was utilized decide the relevancy of the news in this study. In other words, one of the daily tasks of researcher's occupation involves deciding relevancy of the news as mentioned earlier. Hence, that was taken as the basis to decide whether a news "Relevant" or "Not-Relevant".

However, only relying on researcher's personal experience would not be adequate. Hence, it is important to justify the annotation based on previous literature. The study of Cuellar-Fernández, Fuertes-Callén and Laínez-Gadea (2010) found evidence that news related to marketing and promotions, awards and prizes, news related to enhancing the firm's reputation or building a positive public image had no effect on stock market decisions. Therefore, this type of news can be considered "Not-Relevant." On the other hand, evidence suggested that news relevant to new product launches, new customers, business alliances, strategic decisions, acquisition, and entering new markets have significant impact on stock market decisions. Accordingly, these types of news can be categorized as "Relevant."

In this study, the annotation was performed manually by labelling each news document. Based on above-mentioned justification, the two news categories were defined as follows. "Relevant" category refers to the news that are more relevant to gain business insights such as business partnerships, product launches, regulation changes etc. On the

other hand, “Not-Relevant” category represents news that are less relevant to business decision making such as corporate social responsibility (CSR) projects, award winnings, statements by politicians etc.

After identifying two categories for annotation, it is essential to ensure the balance between two data sources with annotating same amount of news for both categories. The study of Padalko *et al*. (2023), and Chen (2024) highlighted the importance of having a balanced dataset to ensure accurate classifications. Accordingly, 500 news articles from Daily FT were assigned to the “Relevant” category and 500 to the “Not-Relevant” category. The same process was applied to the news from Lanka Business Online too. This is demonstrated in Figure 3.3 This ensures facilitating unbiased performance of classifier by treating two classes as well as two news sources equally.

Table 3.2 Composition of Datasets

Dataset	Total number	Daily FT	Lanka Business Online
Mini-dataset (To train the classifier)	2,000	1,000	1,000
Main-dataset	6,000	3,000	3,000
Total news documents	8,000	4,000	4,000

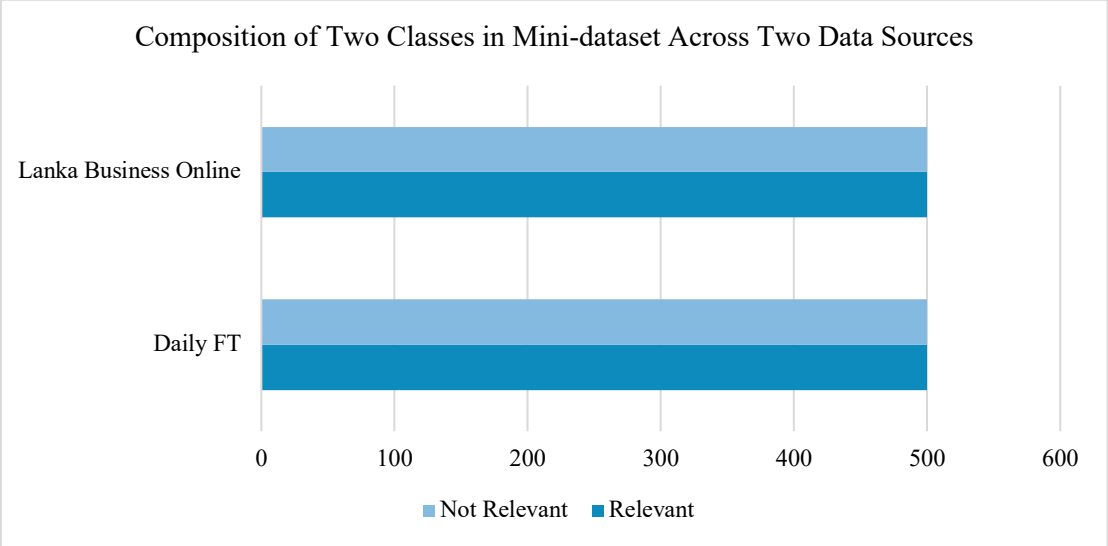


Figure 3.3 Composition of number news articles across two classes in mini dataset

3.2.7 Data Pre-processing

The scraped data was in raw format. However, to perform algorithms on top of raw data requires transforming them into a suitable form. Hence, it was required to perform necessary preprocessing tasks such as removing html tags, tokenization, correcting unicode errors, and removing special characters to ensure that the data is ready for subsequent processes. Since the research is combination of multiple methods such as document classification, document clustering, it required to perform these tasks separately for each task. For instance, tokenization needs to be performed to train the classification, perform the classification on new data, and to perform clustering as well. Each pre-processing tasks carried under those stages are explained under each task later, while a broader explanation is provided here about pre-processing.

a) Combining Columns

The scraped dataset included columns for news date, news header, news content, and news link. However, for the purpose of research, only the news headline and content were required. Thus, first step was to concatenate those two into a single column (Daud *et al.*, 2023). Song (2022) suggested that initial part of the initial part of the news content often contain most important information. Additionally, Chen (2024) conducted a news classification on Chinese headlines, emphasizing the significance of the news header for text analytics tasks. Hence, combining news header with news content would enrich textual data by allowing machine learning models to gain more contextual information.

b) Correcting Unicode and Other Errors

After combining the headline with content, unicode errors were another problem which needed to be addressed prior to performing all text analytics tasks. The data was encoded utilizing 'ISO-8859-1' to prevent the error "*UnicodeDecodeError: 'utf-8' codec can't decode byte 0xd5 in position 366: invalid continuation byte*," which occurred when loading the dataset to Python environment (here Jupyter Notebook). However, still there were still mis-encoded characters such as √**ï,** ¬. Further, there were cases which includes URLs (e.g. http://). Additionally, some news includes the news date within the content. Those three issues were solved by utilizing customized functions to remove those mis-encoded characters, URLs, and dates. Removal of these ensure efficiency of the text analytics tasks as it removes the unnecessary tokens.

3.2.8 Classification of News

a) Building the Classifier with Mini-dataset

The mini dataset was utilized to build the classifier. The reason for employing a classifier was to automate the news selection task and avoiding manual removing not-relevant news documents. Thus, the purpose was to build a classifier to select "Relevant" news articles and omit "Not-relevant" ones. The study employed BERT model which is a deep learning-based technique for classification task. As the name suggests the model employs an encoder architecture, making it more suitable for NLU related tasks such as text classification. BERT is a pre-trained model developed by Google. Since BERT is already trained on large datasets, it did not require to train it from the scratch. However, it required to fine-tune for the specific task, which in this case involves classifying news into two categories.

Since, BERT is a neural network-based classifier, it did not require feature extraction as BERT automatically learn features. BERT has several variants, including BERT-Tiny, BERT-Mini, BERT-Small, BERT-Medium, BERT-Base, and more. Song (2022) utilized different versions of BERT models, including NB-BERT-large, a model built on the large digital collection at the National Library of Norway. The study concluded that BERT models were an appropriate method for news text classification, with NB-BERT-Large achieving the highest accuracy. However, this was for textual data in the Norwegian language and not in English, which may affect the generalizability of the results to English-language datasets. However, the study by Nugroho, Sukmadewa, and Yudistira (2021) demonstrated that the BERT-Base model achieved higher accuracy for news articles in English. Therefore, this study adopts the BERT-Base model. Further, Chen (2024) also suggested similar findings regarding the effectiveness of BERT for news classification, which further justifies its usage in this study.

As mentioned previously, dataset containing 2,000 news documents (mini-dataset) were employed to train the classifier. Initially, data pre-processing methods specified in section 3.2.7 was applied to clean the dataset. Transformers library was utilized to import BertModel. To implement the model PyTorch library was utilized. Further, BertTokenizer in the transformers library was employed to tokenize the text data. Additionally, train-test split of 70:30 (Chen, 2024), and 80:20 (Song, 2022) were experimented utilizing sklearn.model_selection libray.

Prior to ingesting the news documents into the BERT model, they were organized in a particular order to mitigate any potential biases. The documents were arranged with one article from DailyFT followed by another from Lanka Business Online, and so on. This arrangement ensured that the model was not biased towards any specific source when determining the train-test split.

The training was experimented over 10 epochs as indicated by Chen (2024). The model was evaluated in terms of accuracy, precision, recall, F1-score, and receiver operator characteristics curve. Even though the study utilized a balanced data set it would be important to calculate other metrics other than accuracy. The reasons are ensuring more detailed understanding on model performance and ensuring comparability of results with previous studies that have utilized different evaluation metrics. The sklearn.metrics library was utilized to obtain those metrics.

b) Perform the Classification on Main-dataset

After training the classifier, it was saved to perform classification on main-dataset. Thereafter, the saved classifier was applied for the main-dataset to classify "Relevant" and "Not-relevant" news documents. In a similar manner to the previous stage, data pre-processing methods specified in section 3.2.7 was applied to clean main-dataset, before performing the classification. The PyTorch library was utilized to implement the classifier, and BertTokenizer was employed for text data tokenization.

Here, it is important to note that, only the news documents classified as "Relevant" were passed into next phase of the process, which is clustering of news. This is because the aim of the research should be supported with valid dataset. Hence, to gain meaningful business insights from online news data, it is essential to utilize most relevant news documents. After performing the classification, the output was saved in a CSV file.

3.2.9 Clustering of News

After the classifying the main-dataset, news clustering was performed by utilizing the K-means algorithm on the news documents that were classified as "Relevant". Some researchers have utilized R programming to perform K-means algorithm (Bano *et al.*, 2018). However, K-means algorithm is available in Python sklearn library (Sarkar, 2019). Thus, study employed sklearn library to perform news document clustering.

Utilizing the K-means algorithm simplifies understanding for users in the business domain, while also ensuring lower computational costs with increasing volume of data. Further, it ensures scalability to utilize different K values based on the domain knowledge of the user. For example, for a particular product, if the business knows there are only 4 categories of customers, based on that domain knowledge, K can be taken as 4. On a similar note, if the user's focus is more towards the stock market, the number of sectors in the stock market can be taken as the K value (Turner, 2021). Accordingly, the scalability of the K means method makes it more suitable to apply in the business domain.

Prior to perform clustering it was required to extract features. The research study utilized TfidfVectorizer from the sklearn.feature_extraction.text library which is a library available in Python (Shevendrakumar, 2023). Further, study of Kumbhar *et al*. (2020), and Adek, Dinata and Ditha (2021) also utilized the same method for feature extraction for document clustering. When applying TfidfVectorizer on the "Relevant" documents, four parameters were introduced. Those are stated below.

a) **min_df** – This refers minimum document frequency. This means the minimum number of documents a word must appear to be included in the feature set. In this study, min_df =5 where the words that appear less than 5 documents were ignored.
b) **max_df** – This the maximum document frequency that denotes the words that appear more than specified threshold is ignored when deciding the feature set. The max_df = 0.95 where the words that appear in more than 95% of the documents were ignored.
c) **max_features** – This denotes the maximum number of features (here words) is kept for applying TfidfVectorizer. There were around 10,000 unique words in the documents that classified as "Relevant". Hence, that value is utilized for this parameter.
d) **stop_words** – Stop words have little or no significance. The words such as a, the, me etc. are stop words. Therefore, it would be appropriate to remove stop words before perform document clustering (Kumbhar *et al*., 2020). The documents in the study are in English, hence, this parameter was set into 'english'.

After extracting features, the clustering task can be performed. However, for K-means, it is required to decide a suitable K value which is the most important parameter. This research utilized the elbow method along with silhouette score to determine the optimal number of clusters, similar to the approaches employed in previous studies by Rusu, Hodson and Kimball (2014), and Marutho *et al*., (2018). Further, Matplotlib plotting library was utilized to plot the clustered data points. At the end of clustering, the document and respective cluster id of the document were obtained and saved in a CSV file. This is input of the next phase of the study which is the abstractive text summarization.

3.2.10 Abstractive Text Summarization

After obtaining the document clusters, it is important to know what each cluster means. In other words, an average businessperson is not be able to comprehend clusters effectively. Thus, a text summarization method is useful to fulfill that gap. However, utilizing extractive summarization would not effective when a cluster contains a large number of news articles. This is because extractive methods extract key phrases from each news article which is a main pain point with higher number of news articles. Hence, similar to study of Keswani *et al*. (2024), utilizing an abstractive method is fruitful as it reduces information overload, and ensures a short but comprehensive summary for each cluster.

a) Large Language Models (LLMs)

For this task, previous research studies have utilized various open-source pre-trained models such as Hugging Face library, BART, BERT, PEGASUS to generate abstractive summaries (Li *et al*., 2023), (Wang *et al*., 2021), (Patel, 2022), and (Tripathi and Pandey, 2024). In fact, all of these can be utilized and picked the best one based on ROUGE score as suggested in study (Moratanch and Chitrakala, 2016). In previous studies Wang *et al*. (2021), and Li *et al*. (2023), had utilized ROUGE score. However, a key concern with utilizing the Hugging Face library was that, despite being free, it could not be utilized at one point due to the usage limits set by the platform. On the other hand, models such as Google PaLM can be employed utilizing an API key; however, the API key expires after a certain period. This made Hugging Face library, and free model that requires API keys not suitable

for experimental nature research as it could limit the usage in the middle of experiment.

This leads to another option of utilizing paid models such as ChatGPT, Claude. However, it would cost some considerable amount of money due to multiple experiments of large text dataset. Hence, a method was required to conduct the experiment without the aforementioned technical and cost limitations. This was achieved by installing Ollama locally on the laptop. Ollama is a tool that allows running LLMs on local machine. Further, it has various pre-trained models such as Meta's Llama, Mistral AI, and DeepSeek with different number of parameters. Since the research utilized a computer with limited processing power, models with fewer parameters were installed. Specifically, the Llama3 model with 8 billion parameters, the Mistral model with 7 billion parameters, and the DeepSeek-R1 model with 7 billion parameters were installed.

b) Framework to Interact with LLMs and Other Tools

Since the study employed LLMs to develop an application for abstractive text summary generation, a framework tool was necessary. Therefore, the study utilized LangChain, a software framework that enables the integration of LLMs into applications which is similar to the study of Manathunga and Illangasekara (2023). Further, dealing with text data requires to employ embedding models to convert textual data to numerical form. Ollama provides "nomic-embed-text," a free text embedder. This was integrated to the application by utilizing LangChain's OllamaEmbeddings.

Further, to read the data, LLMs requires the data to be stored in a vector database which is a database to store and manage data as high-dimensional vectors. For this purpose, "Chroma", open-source AI application database from langchain.vectorstores was utilized. Data is stored in a vector database along with its associated metadata. Since the goal of the research is to generate summaries for clusters, cluster id was stored as metadata. Shevendrakumar (2023) suggested that utilizing metadata could help gain more insights about data. To store data in vector database, it required to chunk the data. Chunking means splitting large textual data into smaller pieces. Similar to Manathunga and Illangasekara (2023), for this

purpose, “RecursiveCharacterTextSplitter” from langchain.text_splitter was utilized with specifying parameters of chunk_size, chunk_overlap.

c) Retrieval Augmented Generation (RAG)

Since the LLMs are trained based on its own dataset, it may not be effective, when utilizing a dataset that it has not seen earlier. The study utilized a specific dataset for its purpose. Therefore, it was necessary to identify a mechanism to incorporate additional knowledge in the specific dataset that may not include in the LLM. Further, the above mentioned pre-trained language models in Ollama were not specifically trained for the task performed in this study. Thus, the capabilities of those language models need to be improved with methods such as RAG similar to Manathunga and Illangasekara (2023), Keswani *et al*. (2024), and Alkhalaf *et al*. (2024). This technique considers the contextual data into context when generating summaries and attempts to minimize the hallucinations. This was evident in the study by Manathunga and Illangasekara (2023), where an LLM integrated with RAG provided specific and accurate answers, whereas the LLM without RAG produced more generic responses.

For that task, LangChain’s “ChatPromptTemplate”, and “RunnablePassthrough” tools were utilized. It creates a chain by integrating the LLM, the prompt, and output parser. For the output parser LangChain’s “StrOutputParser” was employed. Simply, chain structure creates a sequence of steps that prepare data, interact with the LLM, and return the output. Additionally, LangChain’s “ChatOllama” option was utilized to specify the LLM (Llama/ Mistral AI/ DeepSeek), and temperature parameter. Temperature parameter is to balance between the creativity and consistency of the generated content.

3.2.11 Graphical User Interface (GUI)

The output of clustering task and summary generation task was presented by utilizing a suitable graphical user interface created by open-source web application creation framework “Streamlit”. There are two parts in the user interface, the side bar and the main content area. The sidebar serves as the part of the interface where users can input relevant information whereas main content area demonstrates graphs, and textual output.

A graph containing elbow plot along with respective silhouette scores is demonstrated on the top of the main content area. Based on that, the user can decide the K-value and enter the K value in sidebar to perform K-means clustering. Then, the clustering data plot is demonstrated in the main content area. Once the clustering is finished, the user can enter cluster id and obtain the summary which is displayed in the main content area.

3.2.12 Architectural Decisions of the Research Project

For whole research project, architecture related decisions and reason behind those decisions can be summarized as follows. The decisions were taken to support the experimental nature of the study.

a) The scraped news articles were stored in a Comma-Separated Values (CSV) format, and hence, the datasets related to the classification stage, and clustering stage were also in CSV format. The CSV format is widely supported, easy to manipulate, and facilitates simple integration with data processing tools.

b) To facilitate text processing, the RecursiveCharacterTextSplitter from LangChain was employed for chunking the data into smaller pieces. This method is effective for breaking down large documents into smaller, semantically meaningful chunks that preserve context.

c) The chunked text data was stored and indexed in a vector database, utilizing Chroma, an open-source vector store available in LangChain. Chroma is open-source, free vector store available in LangChain, providing a cost-effective solution for efficient semantic search and retrieval, in contrast to paid alternatives.

d) For text embedding, the OllamaEmbeddings model in LangChain was employed, enabling the conversion of textual content into numerical vector representations for efficient semantic search. OllamaEmbeddings is a free and open-source embedding model makes it more suitable for experiment based research, compared to paid alternatives.

e) The response generation pipeline incorporated StrOutputParser**,** ChatOllama**,** ChatPromptTemplate**,** and RunnablePassthrough from LangChain, ensuring structured input processing, prompt formulation, and coherent response generation. The above mentioned are open-source, free tools that provide a cost-effective solution for experimentation compared to paid alternatives.

The architecture diagram for the research is demonstrated in Figure 3.4 below. It demonstrates classification, clustering, and summarization phases of the research study along with the way it is presented to the user to interact.

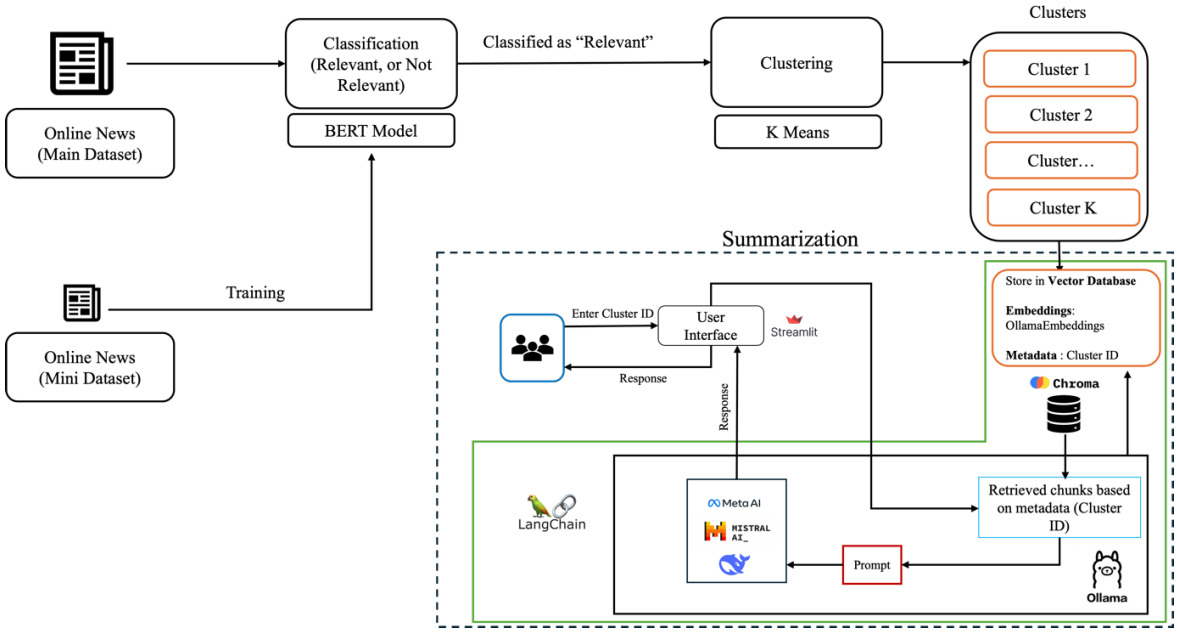


Figure 3.4 Architecture diagram of the research project

3.3 Limitations

The methodology of the study has some limitations too. The number of news articles utilized in the study has to be limited due to the low processing power of the computer utilized. Some previous studies such as Wang *et al.* (2021) had utilized more than 70,000 news documents which could generate more enriched business insights. However, the dataset size of this study was limited to 6,000 news documents. Further, lower processing power made the researcher to utilize LLMs with lower parameters (8 billion maximum). Additionally, the RAG method employed in the study can further be improved by introducing different embedding method and introducing different chunking method such as parent-child chunking.

3.4 Concluding Summary

This chapter outlines the methodology that was employed to achieve the set research objectives. Accordingly, the research employed pragmatic research philosophy with the intention of brining more practical solution to handle the research problem. Further, the research methodology utilized in this research is inductive approach. Employing the archival research strategy, the research utilized longitudinal data which was selected by employing purposive sampling method. The technical tasks of the methodology included news classification, news clustering, and abstractive summary generation. Further, the results of the study were presented, and visualized in a graphical user interface.

CHAPTER 4
EVALUATION AND RESULTS

4.1 Introduction

In any research project, assessment of the project is an important aspect. Initially the researcher set research objectives, followed by proposing a methodology that is the systematic approach to achieve set research objectives. Thereafter, the methodology is applied on a selected dataset to get the results. However, it is important to conduct an evaluation for the study to ensure quality and relevance of the research findings. Further, evaluation is important identify implications for future studies or practical applications. Hence, the objective of this chapter is to provide results of the research study with its evaluation. The research employed experiment-based method for evaluation where the researcher tests different approaches to assess their effectiveness. The study included three phases, document classification, document clustering, and abstractive text summarization. Hence, the evaluation is conducted for three phases separately.

4.2 Evaluation of BERT Classification Model

A classification model can be trained by adjusting different parameters to find most effective model to do the classification. Accordingly, the selected BERT model was experimented by utilizing different parameters such as learning rate, and length of the text. Further, different train-test split was also employed to find a better model. First, the experiment was conducted based on 70:30 train-test split based on Chen (2024), and 80:20 train-test split based on Song (2022) (Table 4.1).

Chen (2024) highlighted the importance of utilizing different learning parameters for enhancing the model's generalization ability. Hence, three learning rates were utilized for the experiment, namely 2e-5 (Song, 2022), 3e-5, and 5e-5 (Chen, 2024). Additionally, BERT model employed in the study can read maximum 512 tokens at a time. However, attention is paid to run the classification with minimum number of tokens to make the model efficient. Hence, different texts lengths less than 512 tokens were experimented, including 128, 192, and 256. This selection is in line with previous studies as the initial part of the news including header gives enough information to perform the classification (Song, 2022). Further, results of Chen (2024)

suggested that utilizing 10 epochs was sufficient for BERT, and hence, model ran for 10 epochs in the current experiment.

4.2.1 Model Selection Evaluation

The initial experiments had the issue of increasing testing loss after several epochs. This was a clear sign of model overfitting. To avoid that, following method was utilized. First all the parameters of BERT model were freeze and unfreeze only final two layers of the model. These are called 'pooler' players (Figure 4.1). This is justified as BERT is already trained and it only required to fine-tune the model, and the computer utilized for the study did not have higher capacity. In this case, model only adjusts parameters final two layers during training.

```
# Freeze the entire BERT base model first
for param in model.bert.parameters():
    param.requires_grad = False

# Unfreeze only the pooler layer
for name, param in model.bert.named_parameters():
    if "pooler" in name:
        param.requires_grad = True
```

Figure 4.1 Parameter adjustment in BERT model

After the above adjustment to the model, the testing loss started declining, indicating no overfitting. For 70:30 split, learning rate 5e-5 and token lengths of 192, and 256 recorded highest accuracy of 0.8150. On the other hand, for 80:20 split, learning rate 5.e-5 and token lengths of 256 recorded highest accuracy of 0.8325. Hence, it was decided as the best model for classification. This accuracy level is higher than the previous studies of Chen (2024) and Padalko *et al.* (2023) where the accuracy was 0.7 and 0.7988 respectively, despite being lower compared to the study of Nugroho, Sukmadewa and Yudistira (2021) where accuracy was 0.9187.

Table 4.1 Results of the Experiment

Learning Rate	Text Length	70:30 Train-Test Split		80:20 Train-Test Split	
		Testing Accuracy	Testing loss	Testing Accuracy	Testing loss
2e-5	128	0.7867	Declining	0.7825	Declining
2e-5	192	0.7933	Declining	0.7850	Declining
2e-5	256	0.7967	Declining	0.8025	Declining
3e-5	128	0.7983	Declining	0.7850	Declining
3e-5	192	0.8033	Declining	0.7975	Declining
3e-5	256	0.8117	Declining	0.8125	Declining
5e-5	128	0.8033	Declining	0.7925	Declining
5e-5	192	0.8150	Declining	0.8200	Declining
5e-5	256	0.8150	Declining	0.8325	Declining

4.2.2 Further Metrics for Evaluation of Selected Model

Since the dataset was balanced, the evaluation can be solely relied on **'accuracy'** metric mentioned above. However, conducting evaluation based on other metrics would be useful to compare results with other studies. Thus, the selected model was further evaluated by utilizing other methods to ensure robustness of the model. Accordingly, the confusion matrix for the model is given below in Figure 4.2. Confusion matrix summaries the results of a classification model. From the 2,000 news documents, 400 news documents (20%) were utilized for testing. From that, 167 news documents that were actually "Not-relevant" has been classified as "Not-relevant", while 166 news documents that were actually "Relevant" has been classified as "Relevant". Only 33 news documents that were actually "Relevant" has been classified as "Not-Relevant", while only 34 news documents that is actually "Not-relevant" has been classified as "Relevant".

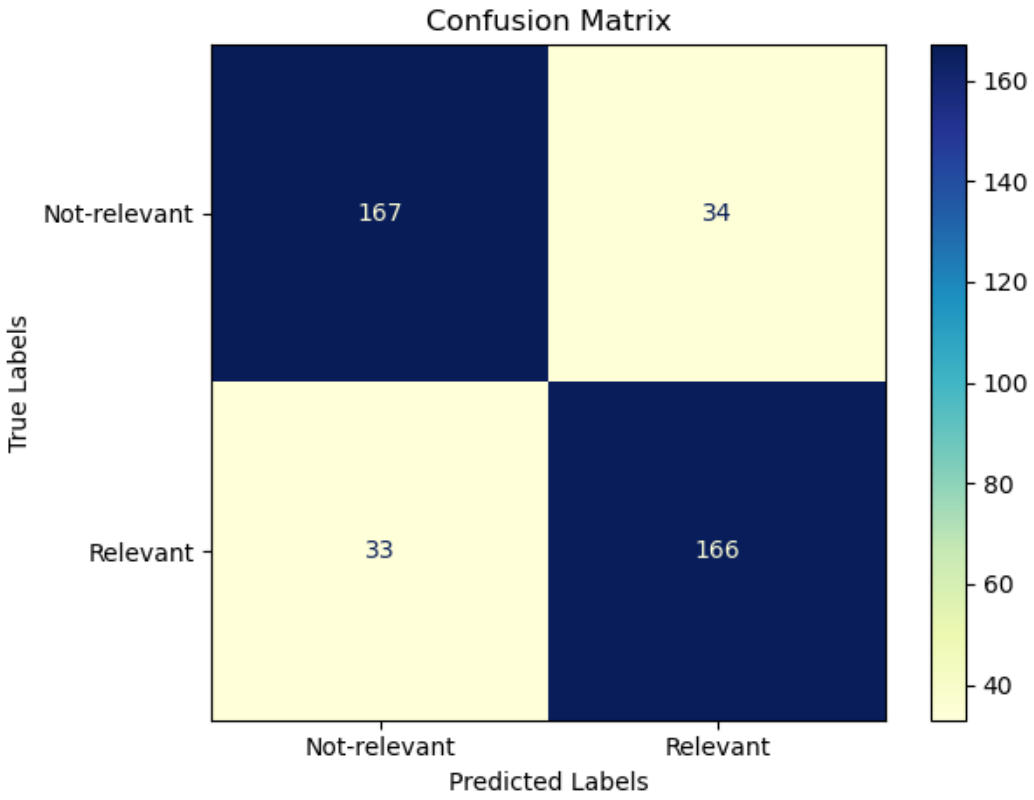


Figure 4.2 Confusion matrix

The confusion matrix not only aids in identifying various classification incidents but also facilitates the calculation of additional metrics to comprehensively evaluate a classification model's performance. Precision, recall, F-1, and accuracy are some common metrics among those. Each metric with respect to the above classification model along with description is given below. Those metrics are presented in the Table 4.2.

Table 4.2 Model Evaluation Metrics

	Precision	**Recall**	**F1-score**	**Support**
Not-relevant (0)	0.83	0.83	0.83	201
Relevant (1)	0.83	0.83	0.83	199
Macro Average	0.83	0.83	0.83	400
Weighted Average	0.83	0.83	0.83	400
Accuracy				0.8325

a) Precision

Precision is the measure of the fraction which a classification model accurately predicts the positive class. In case of this study, precision measures the fraction of cases that the model accurately predicts for “Not-relevant” category, or the fraction of cases model accurately model accurately predicts for “Relevant” category. For the BERT model in the study records a precision value of 0.83, indicating the suitability of the model for a classification task for both classes. This further indicates balanced results for both categories in contrast to cases where the model demonstrates imbalance by predicting one class significantly better than the other, as observed in the study by Padalko *et al.* (2023).

b) Recall

Recall is the measure of the proportion of actual positives that were correctly identified as positives by the model. In case of this study, recall measures the proportion of actual “Not-relevant” cases that were correctly identified as “Not-relevant” by the model or the proportion of actual “Relevant” cases that were correctly identified as “Relevant”. For the BERT classifier in the study records a recall value of 0.83, indicating higher reliability of the model with balanced results for both categories.

c) F1-score

F1-score is a measurement that combines precision and recall. In fact, it is the harmonic mean of precision and recall. The F1-score for the BERT classifier is 0.83, implying the same reliability suggested by previous metrics. Hence, it can be concluded that selected model performs well for classification of both “Not-relevant” and “Relevant” classes.

d) Testing Accuracy over Epochs

It is important to observe the behavior of the testing accuracy over epochs (Chen 2024). There is a significant increase testing accuracy of the model during first three epochs. However, the testing accuracy fluctuates between epoch 3 to 6, followed by slightly stable behavior around the accuracy level of 0.83 over the epochs between 6 to 10. This behavior is similar to the results of Chen (2024) where the testing accuracy stabilized at final epochs. The behavior of testing accuracy is demonstrated in Figure 4.3.

e) Testing Loss over Epochs

The behavior of the testing loss is an important indication to reveal the overfitting of the classification model. The increase in testing loss is a clear indication that a model is suffering from overfitting, implying model's inability to generalization. However, the BERT model in this study demonstrates continuous decline in testing loss similar to the findings of Chen (2024). This is demonstrated in Figure 4.4 below. Hence, the model is good enough for generalization.

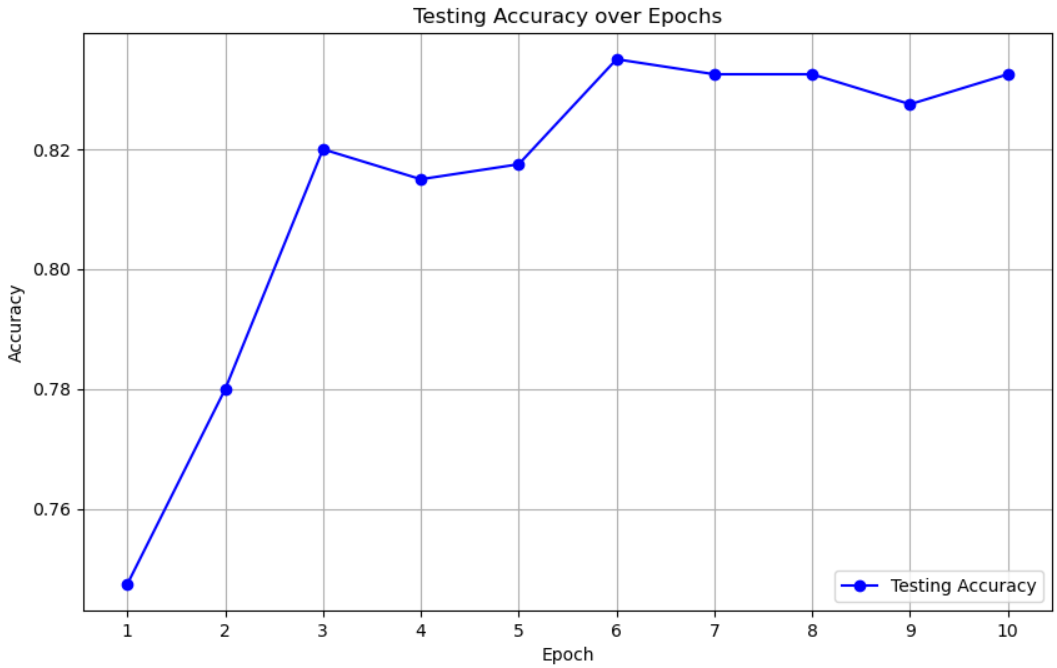


Figure 4.3 Testing accuracy over epochs

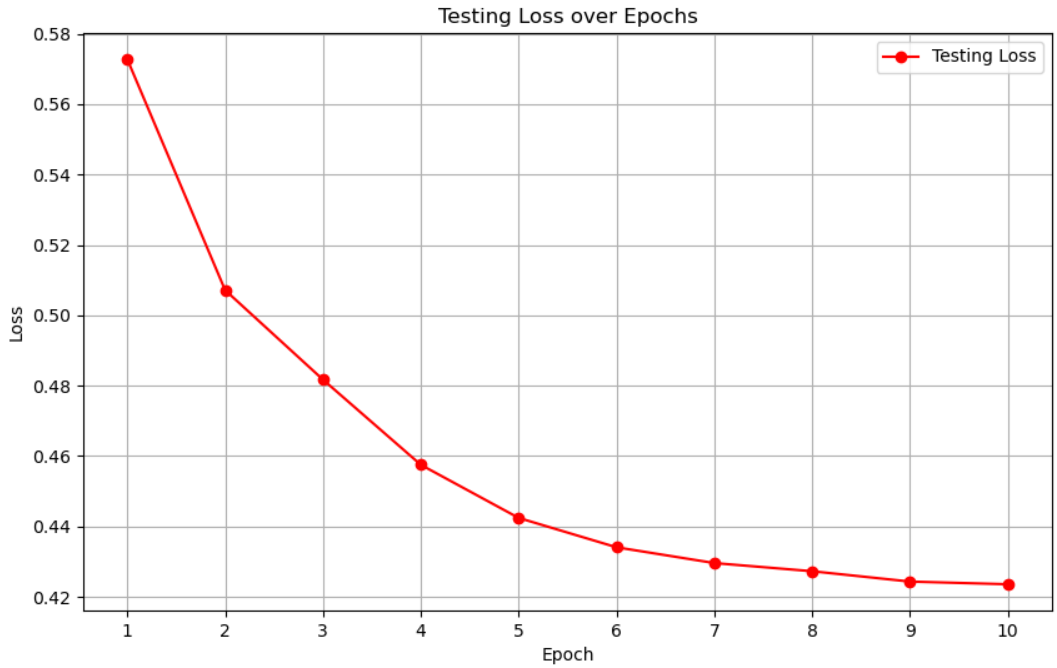


Figure 4.4 Testing loss over epochs

f) Receiver Operator Characteristic (ROC) Curve

Apart from the above measures, ROC curve is graphical method for evaluating a classifier. It plots true positives against the false positives from the confusion matrix. As demonstrated in the Figure 4.5, the dotted line (diagonal line) represents a ransom guess. Hence, it is where accuracy is 0.5. If a model fails to achieve at least this level of performance, it can be concluded that the model is ineffective and inadequate for practical applications. For this study, the ROC curve for the BERT model lies above the random guess line, and hence it can be concluded that the model is effective and adequate for practical applications. However, it is important to integrate a numerical figure to the ROC curve and area under curve (AUC) of ROC curve is a commonly utilized measurement. Accordingly, the BERT model AUC value is 0.83, indicating good level of model performance. Nevertheless, this value is less compared to the previous studies of Song (2022), and Padalko *et al*. (2023) where the AUC values were 0.9492, and 0.87 respectively.

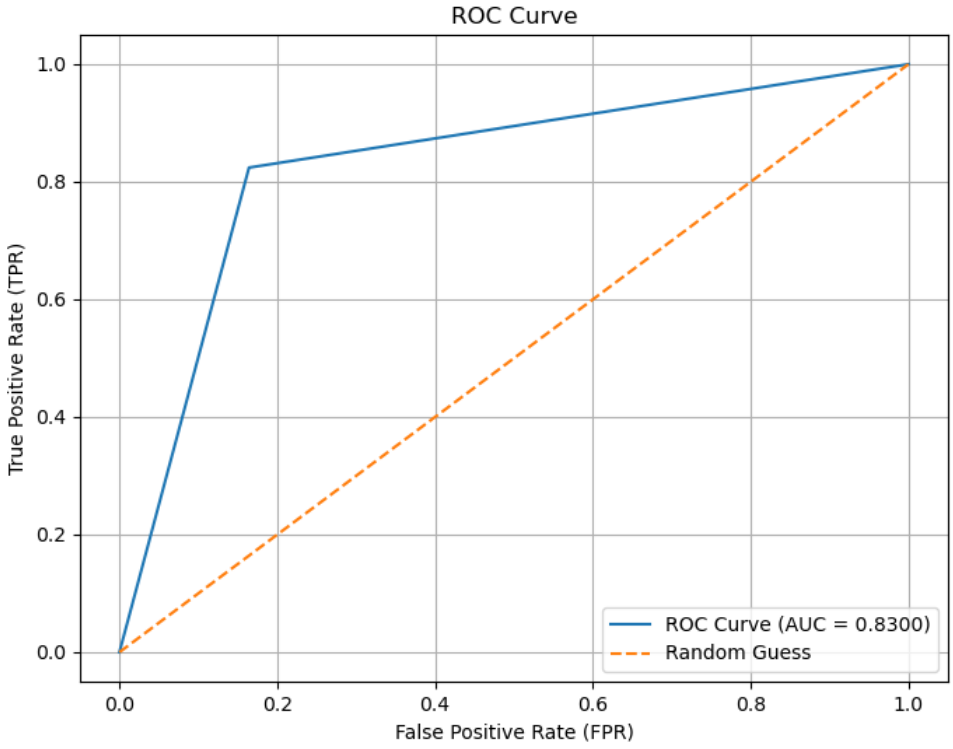


Figure 4.5 Receiver operator characteristics curve

The above classification was applied on main dataset which contained 6,000 news documents. From that only 2,828 news documents were classified as "Relevant" which was the input for second phase of the study, clustering of news documents.

4.3 Evaluation of K-Means Clustering

The next phase of the study is the clustering of documents which were classified as "Relevant". K-Means clustering was employed in the study. The main challenge of utilizing K-Means clustering algorithm is determine a proper K value. Elbow method is one way of deciding a K value. It's a graphical method of finding K value. The graph demonstrates within-cluster-sum-of-square (WCSS) value in Y axis against the corresponding K values in X axis. WCSS is sum of the square of the distances between each data point and its centroid within cluster. Then optimal number for K value is decided based on elbow point of the graph.

For the 2,828 news documents that was classified as "Relevant", WCSS was calculated for a random K values ranging from 2 to 35. Hower, the generated graph demonstrated multiple elbow points making it difficult to decide a K value purely on elbow graph. Accordingly. K

values of 3, 9, 17, 21, 28, and 32 were identified as elbow points as demonstrably circles in Figure 4.6 below.

As a solution for the above difficultly, silhouette score was calculated for each K value. silhouette score is a measurement of the quality of clusters. It is calculated based on "Mean intra-cluster distance" (mean distance between a data point and all other data points in the same cluster), and "Mean nearest-cluster distance" (mean distance between a data point and all other data points of the next nearest cluster). The value of silhouette score ranges from -1 to +1. Higher value indicates well-matched data points within the cluster, and poorly matched to neighboring clusters.

Accordingly, silhouette scores were calculated for K values ranging from 2 to 35. Then for the identified elbow points above, each silhouette score was compared as demonstrated by broken lines in the Figure 4.6. The silhouette score for K = 28 recorded the highest (green broken line) with value of 0.0314 followed by K = 32 (0.0296), K = 21 (0.0276), K = 17 (0.0225), K = 9 (0.0217) and K = 3 (0.0116). Hence, K = 28 was taken as the appropriate K value for the clustering task in the study.

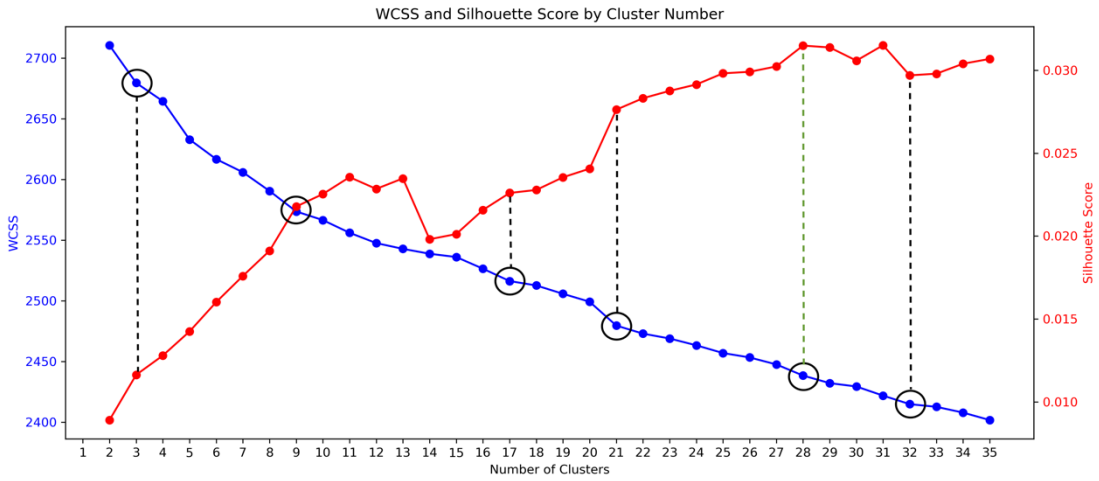


Figure 4.6 Determining K value based on elbow method and silhouette score

Based on the decided K value above, clustering was performed for 2,828 news documents. The data plot before performing the clustering is demonstrated in Figure 4.7 below. The clustered graph is demonstrated in Figure 4.8. After performing the clustering further measurements can also be employed to evaluate the quality of the clusters. Those are homogeneity, completeness, adjusted rand-index, and v-measure. Kumbhar *et al.* (2020) utilized homogeneity, completeness, adjusted rand-index to evaluate K-means document clustering task.

a) Homogeneity

Homogeneity refers to the condition where each cluster contains exclusively data points that belong to a single class (Aljarah *et al*., 2021). Perfect homogeneity is indicated by homogeneity value of 1.0. For the above clustering, homogeneity value of 0.538 ± 0.015. It means on average in each cluster, about 53.8% (0.538 ± 0.015) of the data points belong to single class, suggesting average cluster quality. However, this value is higher than the homogeneity score of Kumbhar *et al*. (2020) where values were 0.248, 0.163, and 0.125 for K values of 2, 4, and 6 respectively.

b) Completeness

Completeness refers to the condition where all data points belonging to a particular class are assigned to the same cluster (Aljarah *et al.,* 2021). Completeness value of 1.0 indicates all data points of a given cluster are perfectly grouped into the same cluster. For the above clustering, completeness value is 0.611 ± 0.021. It means on average 61.1% of the data points belonging to a particular class are correctly assigned to the same cluster. This value is noticeably higher than the completeness score of Kumbhar *et al.* (2020) where values were 0.331, 0.171, and 0.163 for K values of 2, 4, and 6 respectively. Further, compared to homogeneity, slightly higher completeness value suggests good level of clustering quality.

c) Adjusted Rand-Index

Adjusted rand-index evaluates the similarity between predicted and actual clustering labels while accounting for chance (Shevendrakumar, 2023). Hence, adjusted rand-index 1.0, indicating perfect agreement between predicted and actual clustering labels. The adjusted rand-index for the above clustering is 0.246 ± 0.012. It indicates lower level of agreement between the predicted clustering labels and the true class labels, after accounting for chance. Hence, it suggests that there is still room for improvement in the quality of the clustering. Nevertheless, the above value notably higher than the completeness scores reported by Kumbhar *et al*. (2020), which were 0.174, 0.095, and 0.037 for K values of 2, 4, and 6, respectively.

d) V-measure

V-measure is a combination of homogeneity and completeness. It measures how well the criteria of homogeneity and completeness are met (Aljarah *et al*., 2021). The V-measure value of 1.0 indicates perfect clustering quality. For the above

clustering task V-measure value is 0.572 ± 0.014. This value suggests that clusters formed are at average level in grouping similar data points and ensuring that all data points of the same class are grouped together, implying possibility of further improving the cluster quality.

The overall cluster quality assessment of the clustering task is demonstrated in the Table 4.3. It suggests that the cluster quality is maintaining an average level which is similar to the findings of (Shevendrakumar, 2023).

Table 4.3 Overall Cluster Quality Assessment

Measurement	Value	Cluster Quality
Homogeneity	0.538	Average
Completeness	0.611	Above Average
Adjusted rand-index	0.246	Low
V-measure	0.572	Average

Figure 4.7 Data plot before clustering

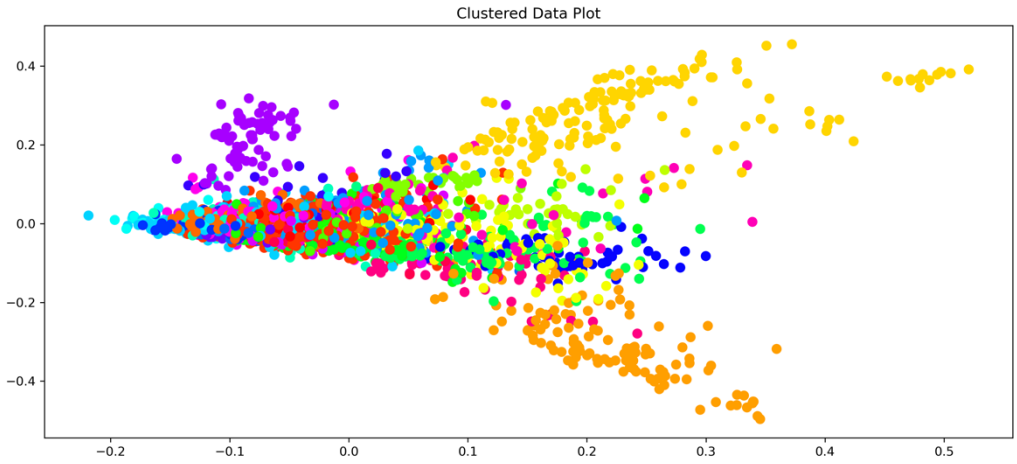


Figure 4.8 Clustered data plot

The breakdown of the number of documents in each cluster from the total 2,828 news documents is demonstrated in Table 4.4. It describes the 28 clusters in terms of types of news in each cluster.

Table 4.4 Description for Each Cluster

Cluster ID	Number of Documents	Description about the Cluster
0	68	News related to the real estate market in Sri Lanka predominantly focuses on the construction of residential properties, investments, and land prices
1	485	Mix of various types of news
2	69	News related to marine services in Sri Lanka predominantly focuses on Sri Lankan ports, shipping companies, and goods arrival through ports.
3	124	News related to Fitch Ratings predominantly focuses on its assignment of credit ratings to Sri Lankan companies
4	167	News related to company profits, and revenues predominantly focuses on annual and quarterly performances.
5	142	Mix of various types of news

(continued)

Table 4.5.Continued

6	37	News related to inflation in Sri Lanka predominantly focuses on changes in the Colombo Consumers Price Index (CCPI), the National Consumer Price Index (NCPI), and inflation rate changes.
7	35	News related to the performance of Sri Lanka's export market predominantly focuses on export value changes and comparisons of values in USD.
8	55	News related to the Colombo Stock Exchange (CSE) predominantly focuses on announcements by the CSE, regulatory changes by the Securities and Exchange Commission (SEC), and trading activities within the CSE.
9	114	News related to the export industry predominantly focuses on initiatives of the Export Development Board, export marketing, the Chamber of Commerce, and the impact of foreign regulations on Sri Lankan exports.
10	148	News related to banks predominantly focuses on online payments, credit cards, payment apps, and new banking products related to online payments.
11	103	News related to the foreign currency exchange market predominantly focuses on USD remittances, restrictions on remittances by the Central Bank of Sri Lanka, currency rates, and foreign currency reserves.
12	16	News related to Samsung Electronics predominantly focuses on their electronic products.
13	72	News related to sustainability initiatives by companies predominantly focuses on carbon emission cuts, carbon footprint verification, green bonds, blue bonds, and other green initiatives.

(continued)

Table 4.6. Continued

14	49	News related to the Hayleys Group predominantly focuses on Hayleys Fentons, Hayleys Fabric, Hayleys Solar, and Hayleys Aventura.
15	113	News related to the energy industry in Sri Lanka predominantly focuses on the Ceylon Electricity Board (CEB), solar power, petroleum, renewable energy, and energy companies.
16	149	News related to company partnerships predominantly focuses on revenue records, acquisitions, business expansions, and investments.
17	72	News related to Sri Lanka's telecommunication industry predominantly focuses on companies such as Dialog, SLT Mobitel, Hutch, and Airtel.
18	49	News related to Colombo Port City predominantly focuses on the Port City Economic Commission.
19	59	News related to the Central Bank of Sri Lanka predominantly focuses on their monetary policy decisions regarding interest rates and their impact on inflation.
20	34	News related to the plantation industry in Sri Lanka predominantly focuses on tea, rubber, and coconut, as well as other factors such as fertilizers, worker wages, crop yields, and overall productivity.
21	51	News related to entrepreneurship predominantly focuses on programs and support for entrepreneurs, startups, startup support, and angel funds.
22	75	News related to the Sri Lankan tea industry predominantly focuses on tea estates, tea factories, tea prices, and tea exports.

(continued)

Table 4.7. Concluded

23	77	News related to Cabinet decisions predominantly focuses on Sri Lankan government projects, new bills, and other legislative matters.
24	297	Mix of various types of news
25	58	News related to the Colombo Stock Exchange predominantly focuses on fundraising, rights issues, debenture issues, and initial public offerings.
26	54	Macroeconomic news predominantly centers on the IMF debt restructuring of Sri Lanka.
27	56	News related to airline services predominantly focuses on Sri Lankan Airlines.

4.4 Evaluation Abstractive Text Summarization

The text summarization employing LLMs was the third part among the three main tasks of the research project. The approach utilized in the study is abstractive text summarization. In the research, different techniques were utilized to increase and measure the quality of the generated summary. Retrieval augmented generation, several LLMs to generate summaries, evaluation metrics, and human evaluation were utilized to evaluate this phase of the research project.

a) Retrieval Augmented Generation (RAG)

RAG is a technique employed to minimize the risk of AI hallucination. By feeding more contextual information for summary generation process, RAG helps improve the summaries generated by LLMs (Keswani *et al*. 2024). The utilization of RAG technique for the summary generation in this study is an important aspect as it helps mitigating risk of generating irrelevant content in the LLM generated summary. Further, prompt is an integral part of RAG. Simply, a prompt is a set of instructions given to the LLM to obtain the intended output by the user.

Prompt

The study employed the prompt template given below. The prompt template consists of assigning a role to the LLM. This is an important method of prompt engineering as assigning a role allows LLM to improve the quality of the output generated. Further, what is expected from the role is also given in the template. Furthermore, instructions

contain facts related to the structure of the output is also given. Additionally, specific instructions are provided to present instances where multiple perspectives or contradictory facts exist. This is mainly because news articles could report same news from different perspectives which is an important supporting factor for gaining business insights to support decision making.

Prompt Template = """

You are a seasoned business and economic advisor, providing expert insights to a strategic decision-making team.

Your role is to extract and summarize valuable business insights from the provided contexts, with relevant examples from the given context.

Output Structure:

- Organize insights under distinct, thematic headings. If insights are related to specific companies, group them under their respective company headings.
- Label the insights sequentially (e.g., Insight 1, Insight 2, etc.) and summarize each with supporting details drawn from the context.
- Where applicable, present conflicting or opposing viewpoints within the facts of the reference text. If no such viewpoints exist, explicitly state that there are no conflicting views or opposing facts.
- If multiple perspectives or contradictory facts are found, ensure they are clearly separated to facilitate effective comparison for informed decision-making.

Context:

<context>

{context}

</context>

Task:

Analyze the context and extract key business insights, noting if there are any opposing views or diverse perspectives.

"""

Chunking Method

In RAG, chunking method is an important step to store data in a vector database. Vector database is a specialized database designed to store and manage data as high-dimensional vectors. To store data in vector database, the input documents need to be divided into chunks. Chunk is a smaller piece of text after splitting a large document into smaller parts. Hence, effective chunking method is important to gain better responses from the LLM. The current study employed "Recursive Character Text Splitter" method for chunking from LangChain that is suitable for long textual data while preserving the contextual content. "Recursive Character Text Splitter" requires specifying a chunk size. The goal of this text splitter is to recursively divide the long text until it reached the specified chunk size. Since the focus of the research is to gain business insights by referring lengthy texts, utilization of this chunking method is more appropriate.

Apart from the chunk size, chunk overlap is also important related to chunking method. The chunk overlap refers to the number of characters that overlap between two consecutive chunks. For most datasets, it is recommended to maintain a chunk overlap ranging from 5% to 20% of the chunk size (Joshi, 2024). Overlapping is important to preserve contextual meaning. Since, the objective of the study is to gain broader level insights, chunk size of 8,000, and chunk overlap of 1,600 (20% of the chunk size).

Temperature Parameter

Temperature parameter is known as the creativity parameter of LLM (Peeperkorn *et al*., 2024). It controls the balance between the creativity and consistency of the generated content. Temperature parameter 0.0–0.4 considered appropriate for factual and precise responses, whereas 0.5–0.7 is considered suitable for general content creation by balancing creativity and consistency. Hence, for the purposes of the current study, two distinct temperature parameters, specifically 0.3 and 0.5, were employed. This ensured that varied output is generated to make a comparable evaluation.

b) Several LLMs to Generate Summaries

The research study utilized "Ollama" which is a tool for running open-source LLMs directly in a computer. In other words, Ollama allows running LLMs locally which

makes it more suitable for experimental research. From the available LLMs in Ollama, current study utilized Meta's Llama3, Mistral AI, and DeepSeek R1. This allows comparison of summaries generated by each LLMs.

c) Evaluation Metrics

As suggested by previous study of Keswani *et al*. (2024), RAGAS framework is a method of evaluating summaries. There are different metrics available in RAGAS including answer relevancy, faithfulness, context recall, context precision by considering taking "question", "ground truth", "answer", "contexts". However, the main issues of obtaining these metrics were the lower processing capability of the computer and not having a ground truth. In this study, the ground truth is a summary created by a human, based on the same reference document provided to the LLM. However, creating such ground truth is not within the scope of this study. Consequently, only the answer relevancy metric was obtained. Answer relevancy refers to the concept that the generated response should directly address the given question (Es *et al*., 2023).

The answer relevancy metrics for each LLM output under two different temperature parameters are demonstrated below in Table 4.5. According to the results, DeepSeek R1 has demonstrated slightly higher answer relevancy value of 0.6165 for temperature parameter 0.3, followed by Mistral (0.6066), and Llama3 (0.6045). However, for the temperature parameter 0.5, Llama3 has demonstrated higher answer relevancy value of 0.6327, followed by Mistral (0.5975), DeepSeek R1 (0.5190). However, it is important to note that overall, the answer relevancy values are at moderate level ranging from 0.5190 to 0.6327. Furthermore, these values are noticeably lower than the answer relevancy values reported in the study by Keswani *et al*. (2024), which achieved a value of 0.93. Further, answer relevancy alone is not adequate to determine the quality of the summaries generated by LLMs. Other metrics are also required along with answer relevancy. Given the technical constraint of limited processing power in the computer to obtain these additional metrics, it is important to carry out alternative types of evaluations, such as human evaluation.

Table 4.8 Answer Relevancy Score for LLMs

LLM	Temperature = 0.3	Temperature = 0.5
Llama3	0.6045	0.6327
Mistral	0.6066	0.5975
DeepSeek R1	0.6165	0.5190

d) Human Evaluation

Human evaluation of AI-generated content is important to ensure quality, as demonstrated by Wang *et al*. (2021), and Li *et al*. (2023). To evaluate the quality of the summaries generated by each LLM, ten individuals were selected for the evaluation process. The previous study by Alkhalaf *et al*. (2024) employed only three human evaluators. However, the current study employed five human evaluators for each summary. These individuals are professionals with a minimum of five years of experience in the fields of investment research, industry research, investment banking, and market research. The ten individuals were divided into two groups, with five individuals in each group. One group is given LLM generated summaries by three models with temperature parameter value of 0.3, and rest is given LLM generated summaries by three models with temperature parameter value of 0.5. Due to the busy nature of these professionals' work, the summaries generated for the cluster with the lowest number of documents were selected. Consequently, cluster 12, which comprised 16 documents, predominantly focusing on news related to Samsung Electronics and their electronic products, was selected for the human evaluation. However, to prevent any pre-existing knowledge about these LLMs from influencing their judgment, the model names were not disclosed to the human evaluators. Summaries generated by each LLM under two different temperature parameters are presented in Appendices section.

The individuals were given a questionnaire containing questions relevant to four criteria suggested by Kryściński *et al*. (2019). Since the research problem is related to gaining business insights, additional aspect related to that was also evaluated. Accordingly, the following aspects were tested in the human evaluation.

a) Fluency: The quality of individual sentences including the summary contains grammatically correct, well-formed sentences with proper punctuation, capitalization, and clarity.

b) **Relevance**: The selection of important content from the source including the summary focuses on the most important points from the source document, excludes unnecessary information, and appropriately highlights key aspects.

c) **Consistency**: The factual alignment between the summary and the source including the summary accurately reflects the facts and numerical data from the source document, maintains consistency without contradictions, and follows a logical flow from one sentence to another.

d) **Coherence**: The collective quality of all sentences including logically connection, smooth flow, well-structured for easy readability, and provides meaningful and sufficient insights.

e) **Insights**: The content generated by the LLM introduce new insights beyond the source document, maintains high overall quality, and accurately follows the instructions provided in the prompt.

For the summaries generated for temperature parameter of 0.3 and 0.5, the evaluations can be summarized based on the above. The results presented in Table 4.6 and 4.7 are based on whether there was a uniform agreement between all evaluators or not.

Table 4.9 Human Evaluation Results for Temperature Parameter 0.3

Criteria	Llama3	Mistral	DeepSeek R1
Fluency	All agreed on accuracy of grammar, however not on other aspects.	All agreed on all aspects.	All agreed on punctuation, however not on other aspects.
Relevance	All agreed that the summary excluded unnecessary information, however not on other aspects.	There was no uniformity in agreement.	There was no uniformity in agreement.
Consistency	There was no uniformity in agreement.	There was no uniformity in agreement.	There was no uniformity in the agreement. (LKR value was added as USD)
Coherence	There was no uniformity in agreement.	There was no uniformity in agreement.	There was no uniformity in agreement.
Insights	There was no uniformity in agreement.	There was no uniformity in agreement.	There was no uniformity in agreement.

The above results indicate that in terms of fluency, Mistral model stands above the other two models. Further, Llma3 model has also been able to generate content with accurate grammar whereas DeepSeek R1 has been accurate with punctuations. On the other hand, in terms of relevancy, Llama3 has been able to exclude unnecessary information in the generated content whereas other models demonstrated poor performance as there was no uniformity in the agreement between all five evaluators.

The performance of three models were poor in terms of consistency, and notably DeepSeek R1 has demonstrated hallucinated currency codes. The evaluation results further indicates that all the three models have demonstrated poor performance in terms of coherence and insights as there was no uniformity in the agreement between all five evaluators.

Table 4.10 Human Evaluation Results for Temperature Parameter 0.5

Criteria	Llama3	Mistral	DeepSeek R1
Fluency	All agreed on accuracy of grammar, however not on other aspects.	All agreed on accuracy of grammar, however not on other aspects.	All agreed on accuracy of grammar, however not on other aspects.
Relevance	There was no uniformity in agreement.	There was no uniformity in agreement.	There was no uniformity in agreement.
Consistency	There was no uniformity in the agreement	There was no uniformity in the agreement	There was no uniformity in the agreement
Coherence	All agreed on smoothness, however not on other aspects	There was no uniformity in agreement.	There was no uniformity in agreement.
Insights	There was no uniformity in agreement.	There was no uniformity in agreement.	There was no uniformity in agreement.

For the content generated with the temperature parameter 0.5, the results indicate that all three models have been able to ensure grammatical accuracy under fluency. Further, Llama3 model has been demonstrated better performance under coherence criteria whereas other models indicated poor performance as per the evaluation. Nevertheless, in terms of relevancy, consistency, and insights criterions all three models have demonstrated poor performance

indicated by not having uniform agreement among evaluators. Overall comments on the evaluators includes specific comments such as the generated content was only focusing on one or two news documents from the given 16 news documents.

Human evaluations indicate that the generated summaries require further improvement. Unlike the findings of Wang *et al*. (2021) and Li *et al*. (2023), where human evaluation confirmed the quality of the generated content, the current study did not yield similar results, highlighting the need for improvements in the summarization approach.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Introduction

After obtaining the results of the research study and evaluating them, the next step is to draw conclusions and provide suggestions for future work. The objective of this chapter is to provide a high-level presentation of the research findings while linking them to the research problem, questions, and the set research objectives. Initially, the overall findings are presented by highlighting indications of the results. Then the discussion moves on to the contribution to the field by the study which is an important aspect for application of the research in real world cases. Thereafter, the chapter includes discussing limitations of the research, lessons learnt, followed by making recommendations for future research.

5.2 Overall Findings

The research was conducted to address the research problem, "How to utilize online business news to gain business insights?". Accordingly, the aim of the research was to propose an approach address the research problem. Text analytics techniques, such as machine learning, are gaining increasing attention due to the growing volume of online textual data, emphasizing the importance of suggesting pragmatic approaches to utilize those data effectively. Accordingly, the overall results of the study suggest that rather than utilizing text analytics techniques in isolation, combining methods such as document classification, document clustering, and abstractive text summarization enhances the ability to gain business insights from large volumes of textual data. Thus, the aim of the research is achieved, as the proposed approach effectively addresses the research problem, despite certain limitations, which are highlighted later in this chapter.

Document classification is an effective method for filtering out textual documents rather than utilizing all at once. This is an important use case, as it enables the automatic filtering of large volumes of textual documents tailored to a particular task such as obtaining business insights. The utilization of pre-trained models such as BERT for document classification is appropriate, as it eliminates the need for feature engineering, and training a model from scratch. Instead, it requires fine-tuning for the particular purpose, making it a cost-effective and time-efficient approach while also reducing the need for extensive expertise in building machine learning

models for real-world business problems. Hence, BERT is a suitable method to classify online business news with higher accuracy. This answers part of the first research question.

While classification enables the filtering of large number of textual documents and the selection of relevant textual documents for a given purpose, it still requires additional mechanisms to uncover hidden patterns within textual documents. Document clustering is an approach that can be applied to group textual documents based on similarity. In fact, document clustering methods allow forming meaningful groups of text documents. Among the document clustering methods, K-means clustering algorithm provides more scalability as well computational efficiency when dealing with large amount of text documents. Hence, K-means clustering is more suitable method in grouping real-word uses cases. However, quality of clustering should be ensured by incorporating multiple methods rather than relying on single metric, such as silhouette score, adjusted rand-index, homogeneity, and completeness alone. This answers the other part of the first research question.

Even though document clustering enables the grouping of similar text documents, the volume of clustered documents may still be overwhelming for human analysis. Reading a large number of documents can be challenging, necessitating further techniques to enhance interpretability. To address this challenge, abstractive summarization is an ideal approach, which has become more feasible with the emergence of LLMs. However, LLMs have limitations, such as hallucinations, which can be minimized by employing techniques such as RAG to enhance accuracy and reliability. Nevertheless, further enhancements for RAG technique are required to obtain improved results. This answers research question 1 and 2.

Human evaluation is an important part of assessing AI generated content. The results of the study indicate that LLMs performed well in terms of fluency, generating content that is grammatically accurate. However, performance was subdued in terms of relevancy, consistency, coherence, and insights related abstractive text summarization. Additionally, hallucinations noticed in some models too. This highlights the importance of keeping human in loop when utilizing LLMs to handle real world problems such as text summarization to gain business insights.

5.3 Contribution to the Field

The research study contributes to both academia and practice by providing valuable insights into the combined application of text analytics techniques for practical use. To the best of our knowledge, the combination of document classification, document clustering, and abstractive text summarization has not been explored before. In fact, most of the previous studies had combined document clustering, and abstractive text summarization only. Further, the combination of both supervised and unsupervised machine learning techniques is another key contribution, as these methods are often applied in isolation in most cases.

On the other hand, the study contributes to the field of business by proposing a pragmatic approach to utilize online business news for gaining business insights, which in turn supports decision-making. The research involved starting with news data scraping, followed by the application of text analytics techniques such as machine learning and retrieval-augmented generation with LLMs, and finished with the development of a graphical user interface. Hence, this is an appropriate supportive tool to be utilized in decision making of business. For example, this could be employed to find business insights about product launches, competitor strategies, regulatory changes etc. Most importantly, this approach saves time by eliminating the need to manually go through large volumes of news data. Additionally, the simple graphical user interface can be easily navigated and understood by an average user, making the tool more practical.

5.4 Limitations

Despite its novelty and contribution to the field, the research also has some limitations. The accuracy of the classification model was at a level that could be improved by training the model on a larger dataset from with incorporating additional sources. Further, the generated clusters demonstrated average-level quality, indicating potential for further improvement. The text summarization task faced several challenges, including the limited processing power of the computer, which made it infeasible to utilize LLMs with large parameters, and run some metrics to evaluate summaries. On the other hand, there were cost considerations as well. High-performance models such as ChatGPT and Claude require a financial expenditure. However, due to the experimental nature of the study, which involved running the models multiple times, paid models could not be employed.

5.5 Lessons Learnt

The whole research process involved applying most of the knowledge acquired during the post graduate degree programme of Master of Business Analytics. It allowed exploring novel methods of solving real-world business problems by utilizing modern text analytics and tools such as machine learning, and LLMs. However, it is important to prioritize data quality above all, as it is the backbone of the project's success. Poor-quality data leads to poor outcomes, even if it is input into high-quality models.

Even though tasks can be automated utilizing machine learning, it is important to keep the human in loop to oversee the whole process and identify bottlenecks. Therefore, evaluation of LLM generated content by human is an important aspect. Nevertheless, it is essential to apply these techniques when and where necessary as it saves time and allows increasing efficiency of business processes.

The combination of text analytics techniques complements each other by leveraging each technique's strengths to solve the problem. For example, while document clustering cannot determine which type of news should be omitted, classification can address that. On the other hand, when handling large amount of text documents, it is useful to identify natural groupings of text documents. Training a model for such classification requires significant data and time, whereas the clustering method offers a more efficient solution for that task. Finally, every research has its strengths as well as limitations. Supporting decisions with proper justification, by referencing prior research and utilizing established techniques, is crucial to establish the credibility of the entire research effort.

5.6 Recommendations for Future Research

Based on the limitations mentioned above, following suggestions can be made to conduct future research. For the current research, news data is scraped from the news. However, different methods, such as utilizing optical character recognition (OCR), can also be employed to extract data stored in other formats such as PDFs, and physical documents. For the classification task, a larger dataset from more than two data sources can be utilized to improve the generalizability of the results. For the clustering task, different feature extraction method such as word embedding can be applied apart from TF-IDF method. The summary generation task can be further improved in several ways. First, by utilizing LLMs with a higher number of parameters (more than 8 billion) than those employed in this study. Second, the focus of the generated

summaries had been limited to one or two news documents, hence there a room for improvement in data retrieval part. Accordingly, enhancing the retrieval-augmented generation process through integrating techniques such as parent-child chunking, re-ranking, query translation, query decomposition is recommended. Additionally, for each cluster, the summary can be integrated with named entity recognition (NER) techniques to identify named entities such as companies, locations, and product names.

5.7 Closing Summary

The chapter discussed the findings of the study by relating them to set research problems, questions, and objectives. Overall findings of the study suggested that text analytics techniques can be utilized to gain business insights from large text corpus. Accordingly, this study contributes to the field mainly as a supportive tool to decision making in a business. However, it is important to employ proper techniques to overcome the limitations of text analytics methods. The combination of text analytics techniques complements solving the research problem well. Finally, suggestions for further improvements have been made for future studies.

APPENDICES

Appendix A

Summaries Generated by Three LLMs with Temperature 0.3. This is the exact formatting of the summary generated by the LLM for the given temperature parameter.

Model : Llama3

Cluster 12 Summary:

Business Insights:
1. **Growing Demand for Foldable Devices**: The analysis highlights a significant increase in demand for foldable devices, with Samsung experiencing a 150% increase in consumers switching to the Galaxy Z Flip3 and a 140% increase compared to the Galaxy S21.

2. **Unique Selling Points**: Both Galaxy Z series devices offer distinct features that appeal to specific customer groups, such as stylish design, portability, innovative folding form factor, and large screen for viewing content and multitasking capabilities.

3. **Collaboration Key to Success**: Samsung's collaboration with Google is crucial in making the foldable experience scalable by providing Android developers with necessary frameworks and resources.

4. **Investment in RD**: Samsung is heavily investing in research and development (RD) to lead the foldable category with even more pioneering technology.

Opposing Views or Diverse Perspectives:

1. **Concerns about Durability**: While the analysis highlights the innovative design of the Hideaway Hinge mechanism, there may be concerns about the durability of the device's folding mechanism over time.

2. **Cost-Effectiveness**: The high demand for foldable devices might lead to increased prices, which could be a barrier for some consumers. Samsung needs to balance innovation with cost-effectiveness to maintain market share.

3. **Competition from Other Brands**: As more brands enter the foldable market, Samsung may face competition in terms of pricing, features, and user experience.

Recommendations:

1. **Continue Investing in RD**: Samsung should continue investing in research and development to stay ahead of the curve and maintain its leadership position in the foldable category.

2. **Diversify Product Lineup**: Samsung could consider expanding its product lineup to cater to different customer segments, such as budget-conscious consumers or those seeking more premium features.

3. **Foster Partnerships with App Developers**: Samsung should continue collaborating with app developers and industry partners to optimize the foldable experience and create new use cases for its devices.

By analyzing these insights, opposing views, and diverse perspectives, Samsung can refine its strategy to capitalize on the growing demand for foldable devices while maintaining a competitive edge in the market.

Model : Mistral

Cluster 12 Summary:

In this context, the key business insights can be summarized as follows:

1. Samsung is launching a new line of foldable smartphones, the Galaxy Z Fold3 and Z Flip3. These devices offer unique features such as an innovative folding form factor, large screens for multitasking, and productivity tools like the S Pen.

2. The company is targeting consumers who value stylish design, portability, and privacy. The ability to use the cover screen without opening the device is a significant selling point for those concerned about privacy and screen time management.

3. Samsung is collaborating with Google to optimize popular apps like Gmail, YouTube, Duo, and Maps for foldable form factors, making the foldable experience more scalable.

4. The launch of these devices has led to a surge in demand for foldable smartphones, with a 150% increase in consumers switching brands for the Galaxy Z Flip3 and a 140% increase compared to the Galaxy S21. This suggests that both devices are giving consumers more incentive than ever to switch to a foldable.

5. Samsung is heavily investing in R&D to lead the foldable category with even more pioneering technology in the future. They are also open to collaborating with app developers and industry partners to drive innovation in the mobile industry.

6. The launch of these devices is part of a broader strategy by Samsung to diversify its product offerings and cater to various consumer segments, including those looking for unique and innovative smartphones.

7. Samsung is planning to release new innovations in the coming years to continue leading the foldable category.

There are no opposing views or diverse perspectives presented in this context. The focus is on the benefits of Samsung's new foldable devices, their collaboration with Google, and their future plans for innovation in the foldable smartphone market.

Model : DeepSeek R1

Key Business Insights from the Context:

1. **Market Positioning and Innovation:**

- Samsung is leveraging its expertise in foldable smartphone technology to capture a growing market segment with stylish and innovative devices like the Galaxy Z Flip3 and Z Fold3.

- These models are positioned as premium alternatives, offering unique features such as water resistance (IPX8 rating), rotating cover screens, and enhanced productivity tools like S Pen support.

2. **Competitive Pricing and Accessibility:**

- The devices are competitively priced, starting at $799 for the 5G version, making them more affordable than many other premium foldable options.

- This affordability is expected to attract a broader audience, including price-sensitive consumers looking for high-quality smartphone alternatives.

3. **Targeted Marketing and Demographics:**

- Samsung's marketing strategy focuses on Gen Z and millennial women, highlighting the devices' stylish design and portability as key drivers for purchase.

- The appeal extends beyond tech enthusiasts to include users seeking innovative features that traditional smartphones lack.

4. **Increased Consumer Demand:**

- There is a significant increase in consumer interest, with sales of the Galaxy Z series devices outperforming previous models by 150% and 140%, respectively.

- This surge indicates strong market acceptance and the effectiveness of Samsung's marketing strategy in driving brand switching.

5. **Strategic Partnerships for Future Growth:**

- Samsung is investing heavily in research and development (RD) to maintain its leadership in foldable technology.

- Collaborations with companies like Google are expected to enhance compatibility with Android apps, expanding the appeal of foldable devices beyond their current ecosystem.

6. **Challenges and Risks:**

- While the features are groundbreaking, there is a risk associated with overvaluing certain capabilities (e.g., water resistance) that may not translate into real-world usage.

- The marketing strategy's effectiveness in resonating with all consumer segments remains a potential challenge, as some consumers may prefer non-foldable devices.

Summary:
Samsung is strategically positioning its foldable devices to capitalize on innovation and targeted marketing. While the technology represents significant advancements, challenges such as market saturation and differing consumer preferences need to be managed effectively for sustained growth.

Appendix B

Summaries Generated by Three LLMs with Temperature 0.5. This is the exact formatting of the summary generated by the LLM for the given temperature parameter.

Model : Llama3

Cluster 12 Summary:

Context Analysis

The provided context is a strategic decision-making team seeking expert insights from a seasoned business and economic advisor on the topic of foldable devices. The text is an in-depth analysis of Samsung's innovative approach to the foldable market, highlighting its pioneering efforts, technological advancements, and consumer demand.

Key Business Insights

1. **Innovation drives growth**: Samsung's relentless pursuit of innovation has led to a significant increase in consumer demand for foldable devices.

2. **Foldable form factor is a game-changer**: The unique design and functionality of foldable devices have created a new category, offering consumers more flexibility and convenience.

3. **Collaboration is key**: Samsung's partnerships with Google and other industry partners have enabled the development of scalable solutions for large-screen and foldable devices.

4. **Consumer demand is driven by features**: The top drivers behind purchases of Galaxy Z Flip3 and Fold3 are stylish design, portability, innovative folding form factor, and features like App Continuity and MultiActive Window.

5. **Investment in RD is crucial**: Samsung's heavy investment in research and development has enabled the company to stay ahead of the curve in the foldable market.

Opposing Views or Diverse Perspectives

None are explicitly mentioned in the provided context. However, it can be inferred that some potential opposing views might include:

1. **Concerns about durability**: Some consumers may be hesitant to adopt foldable devices due to concerns about their durability and ability to withstand repeated use.

2. **Competition from other manufacturers**: Other smartphone manufacturers may be developing their own foldable devices, which could potentially erode Samsung's market share.

Recommendations

Based on the provided context, I would recommend that the strategic decision-making team:

1. **Continue investing in RD**: To stay ahead of the curve and continue innovating in the foldable space.

2. **Foster partnerships**: With other industry players to develop scalable solutions and expand the market for foldable devices.

3. **Focus on features**: That drive consumer demand, such as App Continuity and Multi Active Window, to differentiate Samsung's products from competitors.

By doing so, the team can continue to capitalize on the growth potential of the foldable device market and maintain a competitive edge in the industry.

Model : Mistral

Cluster 12 Summary:

In the provided context, several key business insights can be extracted:

1. Samsung is launching a new line of foldable smartphones, specifically mentioning the Galaxy Z Fold3 and Z Flip3. These devices offer unique features such as an innovative folding form factor, large screens, multitasking capabilities, and productivity features like the S Pen.

2. The devices are designed to cater to different customer groups: those who value style, portability, privacy, and screen time management (Galaxy Z Flip3) and those who prefer a large screen for content viewing, multitasking, and productivity (Galaxy Z Fold3).

3. Samsung is investing heavily in R&D to lead the foldable category and collaborating with app developers and industry partners such as Google to make the foldable experience scalable and optimize popular apps for foldable form factors.

4. The devices have received positive responses, with a 150% increase in consumers switching brands for the Galaxy Z Flip3 and a 140% increase compared to the Galaxy S21. This suggests that the unique features of these foldable devices are appealing to customers and driving demand.

5. Samsung is targeting markets where there is growing interest in foldable devices, indicating that consumer awareness and potential for growth in this category is high.

6. The devices have received an IPX8 water resistance rating, setting a new milestone for foldables. However, it's not mentioned if other brands have achieved the same rating.

7. Samsung is working with partners and developers to drive future innovations in the mobile industry, aiming to steer its growth and evolution.

There are no apparent opposing views or diverse perspectives in this context. The focus is on showcasing the unique features, benefits, and advancements of Samsung's new foldable devices and their potential impact on the market.

Model : DeepSeek R1

Analysis of Key Business Insights from Samsung's Foldable Smartphone Strategy

1. **Market Expansion and Innovation Leadership**
   - Samsung is expanding its product portfolio by introducing innovative foldable smartphones like the Galaxy Z Flip3 and Z Fold3, which feature cutting-edge technologies such as water resistance (IPX8), improved displays, and enhanced productivity tools.
   - The company aims to lead the foldable smartphone market through continuous innovation and strategic partnerships with tech partners.

2. **Customer Appeal and Diverse Features**
   - Foldables are attracting customers valuing style, portability, and innovative design over solely relying on technological advancements.
   - Specific features like privacy controls in cover screens, S Pen support, and large-screen multitasking capabilities drive purchasing decisions for different customer segments.

3. **Competitive Landscape and Consumer Flexibility**
   - While Samsung is the leader in foldable innovation, there's an implication that consumers might switch brands if these devices meet their needs better than alternatives from competitors like Apple or Oppo.
   - The high cost of advanced features (e.g., IPX8 rating) may limit accessibility for some customers.

4. **Research and Development Focus**
   - Samsung is heavily investing in R&D to further innovate foldable devices, including advancements in app compatibility with foldable form factors, as highlighted by collaborations with Google.
   - This strategic focus aims to enhance user experience on foldable devices, potentially overcoming technical barriers but requiring significant resource investment.

5. **Collaborations and App Optimization**
   - Samsung collaborates with tech partners like Google to optimize apps for foldable devices, ensuring seamless functionality across large screens and innovative forms, which could be seen as a strategic effort to overcome potential limitations of foldable technology.

Conclusion:
Samsung's strategy revolves around leveraging innovation in foldable smartphones to meet evolving consumer demands. By focusing on unique features, expanding product lines, and

collaborating with industry partners, Samsung aims to establish itself as the market leader while addressing the diverse needs and preferences of its customers.

Appendix C

Python Scripts

```
import pandas as pd
import matplotlib.pyplot as plt
import os
import torch
from torch import nn
from torch.utils.data import DataLoader, Dataset
from transformers import BertTokenizer, BertModel, AdamW, get_linear_schedule_with_warmup
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report,roc_auc_score, roc_curve

class TextClassificationDataset(Dataset):
    def __init__(self, texts, labels, tokenizer, max_length):
        self.texts = texts
        self.labels = labels
        self.tokenizer = tokenizer
        self.max_length = max_length

    def __len__(self):
        return len(self.texts)

    def __getitem__(self, idx):
        text = self.texts[idx]
        label = self.labels[idx]
        encoding = self.tokenizer(text, return_tensors='pt', max_length=self.max_length, padding='max_length
        return {'input_ids': encoding['input_ids'].flatten(), 'attention_mask': encoding['attention_mask'].f
```

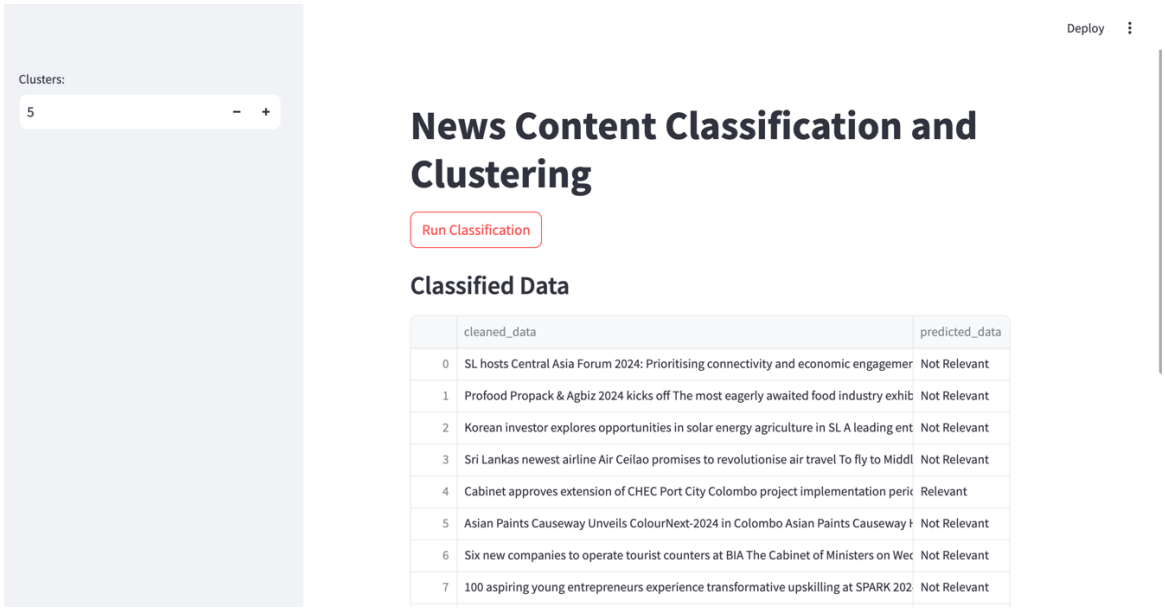
```
# Define the BERTClassifier
class BERTClassifier(nn.Module):
    def __init__(self, bert_model_name, num_classes):
        super(BERTClassifier, self).__init__()
        self.bert = BertModel.from_pretrained(bert_model_name)
        self.dropout = nn.Dropout(0.1)
        self.fc = nn.Linear(self.bert.config.hidden_size, num_classes)

    def forward(self, input_ids, attention_mask):
        outputs = self.bert(input_ids=input_ids, attention_mask=attention_mask)
        pooled_output = outputs.pooler_output
        x = self.dropout(pooled_output)
        logits = self.fc(x)
        return logits
```

```python
import pandas as pd
from langchain.schema import Document
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.embeddings.ollama import OllamaEmbeddings
from langchain.vectorstores import Chroma
from langchain_ollama import ChatOllama
from langchain_core.output_parsers import StrOutputParser
from langchain_core.prompts import ChatPromptTemplate
from langchain_core.runnables import RunnablePassthrough


text_splitter = RecursiveCharacterTextSplitter(chunk_size=8000, chunk_overlap=1600,length_function=len,is_separator_


# Group data by 'Cluster' and process each group
all_splits = []
cluster_docs_map = {}

for cluster_id, group in result.groupby('Cluster'):
    cluster_text = group['cleaned_data'].astype(str).str.cat(sep=" \n ")  # Separate by newlines
    cluster_doc = Document(page_content=cluster_text, metadata={"Cluster": cluster_id})
    cluster_splits = text_splitter.split_documents([cluster_doc])
    all_splits.extend(cluster_splits)
    cluster_docs_map[cluster_id] = cluster_splits  # Store splits by cluster
```

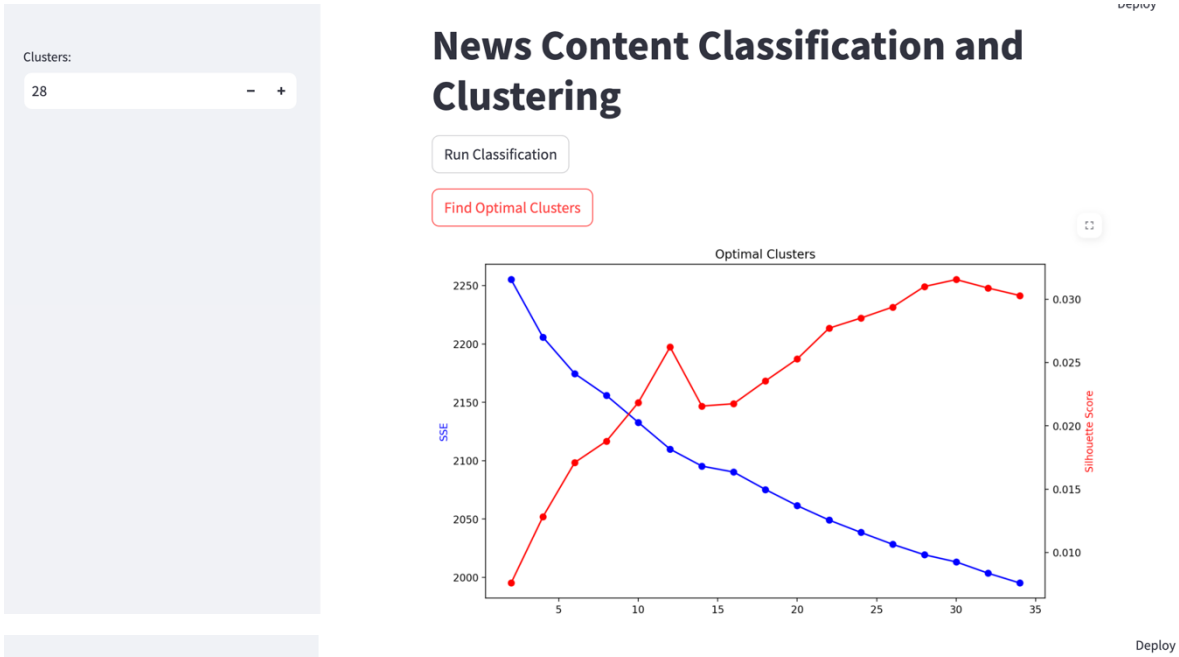
```python
# Initialize the embeddings and vectorstore
local_embeddings = OllamaEmbeddings(model="nomic-embed-text")
vectorstore = Chroma.from_documents(documents=all_splits, embedding=local_embeddings)

# Define the prompt template using the ChatPromptTemplate
rag_prompt = ChatPromptTemplate.from_template(RAG_TEMPLATE)


# Initialize the model for summarization
model = ChatOllama(model="llama3", temperature=0.3)

# Summarization chain
def format_docs(docs):
    return "\n\n".join(doc.page_content for doc in docs)

chain = (
    RunnablePassthrough.assign(context=lambda input: format_docs(input["context"]))
    | rag_prompt
    | model
    | StrOutputParser()
)
```

Appendix D

Streamlit GUI



	cleaned_data	predicted_data
0	SL hosts Central Asia Forum 2024: Prioritising connectivity and economic engagemer	Not Relevant
1	Profood Propack & Agbiz 2024 kicks off The most eagerly awaited food industry exhib	Not Relevant
2	Korean investor explores opportunities in solar energy agriculture in SL A leading ent	Not Relevant
3	Sri Lankas newest airline Air Ceilao promises to revolutionise air travel To fly to Middl	Not Relevant
4	Cabinet approves extension of CHEC Port City Colombo project implementation peric	Relevant
5	Asian Paints Causeway Unveils ColourNext-2024 in Colombo Asian Paints Causeway I	Not Relevant
6	Six new companies to operate tourist counters at BIA The Cabinet of Ministers on Wec	Not Relevant
7	100 aspiring young entrepreneurs experience transformative upskilling at SPARK 202	Not Relevant



REFERENCES

Adek, R.T., Dinata, R.K. and Ditha, A. (2021) 'Online Newspaper Clustering in Aceh using the Agglomerative Hierarchical Clustering Method,' *International Journal of Engineering Science and Information Technology*, 2(1), pp. 70–75. https://doi.org/10.52088/ijesty.v2i1.206.

Ahmed, J. and Ahmed, M. (2021) 'ONLINE NEWS CLASSIFICATION USING MACHINE LEARNING TECHNIQUES,' *IIUM Engineering Journal*, 22(2), pp. 210–225. https://doi.org/10.31436/iiumej.v22i2.1662.

Aljarah, I. *et al.* (2021) 'A comprehensive review of evaluation and fitness measures for evolutionary data clustering,' *Algorithms for Intelligent Systems*, pp. 23–71. https://doi.org/10.1007/978-981-33-4191-3_2.

Alkhalaf, M. *et al.* (2024) 'Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records,' *Journal of Biomedical Informatics*, 156, p. 104662. https://doi.org/10.1016/j.jbi.2024.104662.

Bano, S. *et al.* (2018) 'Document summarization using clustering and text analysis,' *International Journal of Engineering & Technology*, 7(2.32), p. 456. https://doi.org/10.14419/ijet.v7i2.32.15740.

Basyal, L. and Sanghvi, M. (2023) 'Text summarization using large language models: a comparative study of MPT-7B-Instruct, Falcon-7b-Instruct, and OpenAI Chat-GPT models,' *arXiv.org* [Preprint]. https://arxiv.org/abs/2310.10449.

Beach, B. and Hanlon, W. (2022) *Historical Newspaper Data: A researcher's guide and toolkit*, *NBER Working Paper Series*. 30135. https://www.nber.org/system/files/working_papers/w30135/w30135.pdf.

Campbell, S. *et al.* (2020) 'Purposive sampling: complex or simple? Research case examples,' *Journal of Research in Nursing*, 25(8), pp. 652–661. https://doi.org/10.1177/1744987120927206.

Chen, Y. (2024) 'A study on news headline classification based on BERT modeling,' in *Advances in computer science research*, pp. 345–355. https://doi.org/10.2991/978-94-6463-540-9_35.

Chou, H.-C. *et al.* (2024) 'Embracing ambiguity and subjectivity using the All-Inclusive Aggregation Rule for evaluating Multi-Label Speech Emotion Recognition Systems,' *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 502–509. https://doi.org/10.1109/slt61566.2024.10832302.

Cuellar-Fernández, B., Fuertes-Callén, Y. and Laínez-Gadea, J.A. (2010) 'The impact of corporate media news on market valuation,' *Journal of Media Economics*, 23(2), pp. 90–110. https://doi.org/10.1080/08997764.2010.485541.

Daud, S. *et al.* (2023) 'Topic Classification of online news articles using optimized machine learning models,' *Computers*, 12(1), p. 16. https://doi.org/10.3390/computers12010016.

Es, S. *et al.* (2023) *RAGAS: Automated Evaluation of Retrieval Augmented Generation*. https://arxiv.org/abs/2309.15217.

Ismail, I.A., Chukwuka, E.E. and Abiola, M.K. (2024) 'IMPROVING BUSINESS INSIGHTS USING MACHINE LEARNING ALGORITHMS,' *GSJ*, 12–12(2), pp. 1609–1610. https://www.globalscientificjournal.com/researchpaper/Improving_Business_Insights_using_Machine_Learning_Algorithms.pdf.

Joshi, A. (2024) *How to choose the right chunking strategy for your LLM application | MongoDB*. https://www.mongodb.com/developer/products/atlas/choosing-chunking-strategy-rag/.

Keswani, G. *et al.* (2023) *View of abstractive long text summarization using large language models*. https://www.ijisae.org/index.php/IJISAE/article/view/4500/3160.

Kopackova, H., Komárková, J. and Sedlak, P. (2008) 'Decision making with textual and spatial information,' *WSEAS TRANSACTIONS on INFORMATION SCIENCE & APPLICATIONS*, Issue 3, p. 258. http://www.wseas.us/e-library/transactions/information/2008/25-497.pdf.

Kumbhar, R. *et al.* (2020) 'Text Document Clustering Using K-means Algorithm with Dimension Reduction Techniques,' *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, pp. 1222–1228. https://doi.org/10.1109/icces48766.2020.9137928.

Li, H. *et al.* (2023) 'Abstractive financial news summarization via Transformer-BILSTM encoder and Graph Attention-Based Decoder,' *IEEE/ACM Transactions on Audio Speech and Language Processing*, 31, pp. 3190–3205. https://doi.org/10.1109/taslp.2023.3304473.

Mallick, C. *et al.* (2018) 'Graph-Based text summarization using modified TextRank,' in *Advances in intelligent systems and computing*, pp. 137–146. https://doi.org/10.1007/978-981-13-0514-6_14.

Manathunga, S.S. and Illangasekara, Y.A. (2023) *Retrieval Augmented Generation and Representative Vector Summarization for large unstructured textual data in Medical Education*. https://arxiv.org/abs/2308.00479.

Marutho, D. *et al.* (2018) 'The determination of cluster number at K-Mean using elbow method and purity evaluation on Headline News,' *2018 International Seminar on Application*

for Technology of Information and Communication, pp. 533–538. https://doi.org/10.1109/isemantic.2018.8549751.

Moratanch, N. and Chitrakala, S. (2016) 'A survey on Abstractive text Summarization,' 1 March. https://doi.org/10.1109/iccpct.2016.7530193.

Nugroho, K.S., Sukmadewa, A.Y. and Yudistira, N. (2021) 'Large-Scale news classification using BERT language model: SPARK NLP approach,' 13 September. https://doi.org/10.1145/3479645.3479658.

Padalko, H. *et al.* (2023) *Misinformation detection in political news using BERT model*. https://ceur-ws.org/Vol-3641/paper11.pdf.

Patel, P.M. (2022) 'Financial News Summarisation using Transformer Neural Network,' *Research Square (Research Square)* [Preprint]. https://doi.org/10.21203/rs.3.rs-2132871/v1.

Peeperkorn, M. *et al.* (2024) *Is temperature the creativity parameter of large language models?* https://arxiv.org/abs/2405.00492.

Roychowdhury, S. *et al.* (2024) *Evaluation of RAG metrics for question answering in the telecom domain*. https://arxiv.org/abs/2407.12873.

Rusu, D., Hodson, J. and Kimball, A. (2014) 'Unsupervised techniques for extracting and clustering complex events in news,' 1 January. https://doi.org/10.3115/v1/w14-2905.

Sarkar, D. (2019) *Text Analytics with Python*, *Apress eBooks*. https://doi.org/10.1007/978-1-4842-4354-1.

Shevendrakumar, D.D. (2023) 'Clustering and Retrieval of News articles using Natural Language Processing,' *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 07(07). https://doi.org/10.55041/ijsrem24154.

Song, L. (2022) *Ensuring Brand Safety by Using Contextual Text Features: A Study of Text Classification with BERT*, *Master's Thesis in Language Technology*. thesis. https://uu.diva-portal.org/smash/get/diva2:1775452/FULLTEXT01.pdf.

Tarannum, S. *et al.* (2021) NLP based Text Summarization Techniques for News Articles: Approaches and Challenges. [online] International Research Journal of Engineering and Technology. IRJET. Available at: https://www.irjet.net/archives/V8/i12/IRJET-V8I12220.pdf [Accessed 10 Mar. 2025].

Tripathi, P. and Pandey, A.K. (2024) *News summarization articles by using NLP*, *IRE Journals*. journal-article. https://www.irejournals.com/formatedpaper/1705416.pdf.

Turner, E. (2021) *Graph Auto-Encoders for financial clustering*. https://arxiv.org/abs/2111.13519.

Wang, J. *et al.* (2021) 'Unsupervised graph-clustering learning framework for financial news summarization,' *2021 International Conference on Data Mining Workshops (ICDMW)*, pp. 719–726. https://doi.org/10.1109/icdmw53433.2021.00094.

Yadav, D., Desai, J. and Yadav, A.K. (2022) *Automatic Text Summarization Methods: A Comprehensive Review*. https://arxiv.org/abs/2204.01849.

Yu, Z. *et al.* (2021) 'News Credibility and Influence within the Financial Markets,' *Journal of Behavioral Finance*, 24(2), pp. 238–257. https://doi.org/10.1080/15427560.2021.1974443.

Zhang, M. *et al.* (2023) 'ROUGE-SEM: Better evaluation of summarization using ROUGE combined with semantics,' *Expert Systems With Applications*, 237, p. 121364. https://doi.org/10.1016/j.eswa.2023.121364.