CLASSIFICATION OF PAST EXAM QUESTIONS BY SUBJECT CHAPTERS BASED ON MACHINE LEARNING

MOUNZIR MOHAMED HASHIM MOHAMED

UNIVERSITI KEBANGSAAN MALAYSIA

CLASSIFICATION OF PAST EXAM QUESTIONS BY SUBJECT CHAPTERS BASED ON MACHINE LEARNING

MOUNZIR MOHAMED HASHIM MOHAMED

PROJECT SUBMITTED IN PARTIAL FULFILMENT FOR THE DEGREE OF
MASTER OF DATA SCIENCE

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

KLASIFIKASI SOALAN PEPERIKSAAN LEPAS MENGIKUT MATA PELAJARAN BAB BERDASARKAN PEMBELAJARAN MESIN

MOUNZIR MOHAMED HASHIM MOHAMED

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPEROLEH IJAZAH
SARJANA SAINS DATA

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

19 July 2024 MOUNZIR MOHAMED HASHIM MOHAMED
P129289

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious, the Most Merciful, I express my deepest gratitude for his guidance and blessings throughout my academic journey.

I am profoundly grateful to my supervisor, Dr. Nor Samsiah Sani, for her exceptional guidance, insightful feedback, and unwavering support. Her mentorship has been invaluable in shaping this research.

I extend my heartfelt thanks to my family for their unwavering support and encouragement, especially during the challenging times our beloved Sudan is facing. To my parents and siblings, your sacrifices, prayers, and belief in me have been my greatest motivation. Despite the hardships brought on by the ongoing conflict, your strength and resilience have been my inspiration.

Thank you all.

ABSTRAK

Peningkatan eksponen dalam jumlah kertas peperiksaan yang lalu telah menimbulkan cabaran besar kepada pendidik dan pelajar dalam mengkategorikan dan mengakses bahan kajian yang berkaitan dengan cekap. Penyelidikan ini menangani masalah klasifikasi manual soalan peperiksaan ke dalam bab yang telah ditetapkan, yang memakan masa dan tidak konsisten. Matlamat utama kajian ini adalah untuk membangunkan model terbaik untuk pengelasan automatik soalan peperiksaan ke dalam bab khusus subjek menggunakan teknik pembelajaran mesin lanjutan, terutamanya dalam Pemprosesan Bahasa Semulajadi (NLP). Memfokuskan pada soalan aneka pilihan IGCSE Physics (0625), kajian ini menilai pelbagai teknik pengekstrakan ciri termasuk Bag-of-Words, N-Grams, TF-IDF dan model benam perkataan untuk menentukan keberkesanannya dalam meningkatkan ketepatan pengelasan. Penyelidikan menggunakan model pembelajaran mesin seperti Logistic Regression, Random Forest dan Convolutional Neural Networks (CNN) dengan pengoptimuman sistematik melalui penalaan hiperparameter untuk mengenal pasti model berprestasi terbaik. Keputusan menunjukkan bahawa CNN menunjukkan ketepatan tertinggi pada 93.3% antara model yang dikaji. Tambahan pula, ujian statistik mengesahkan perbezaan ketara antara model, dengan CNN mengatasi prestasi kedua-dua LR-TFIDF dan RF-BoW. Model yang dicadangkan memperkemas proses penyediaan peperiksaan dan menyediakan pendekatan piawai untuk pendidik dan pelajar, memudahkan pemahaman yang lebih pantas tentang prestasi dan strategi belajar yang lebih berkesan. Penyelidikan ini menyumbang kepada bidang ini dengan memperkenalkan sistem klasifikasi automatik yang cekap dan boleh dipercayai untuk soalan peperiksaan, yang boleh disepadukan ke dalam pelbagai platform pendidikan, dengan itu menambah baik amalan pendidikan dengan mengurangkan beban kerja pendidik dan menyediakan pelajar dengan bahan pembelajaran yang tepat pada masanya dan tepat.

ABSTRACT

The exponential increase in past exam papers has posed significant challenges for educators and students in efficiently categorising and accessing relevant study materials. This research addresses the problem of the manual classification of exam questions into pre-defined chapters, which is both time-consuming and inconsistent. The primary aim of this study is to develop the best model for the automated classification of exam questions into subject-specific chapters using advanced machine learning techniques, particularly in Natural Language Processing (NLP). Focusing on IGCSE Physics (0625) multiple-choice questions, this study evaluates various feature extraction techniques, including Bag-of-Words, N-Grams, TF-IDF, and word embedding models, to determine their effectiveness in enhancing classification accuracy. The research employs machine learning models such as Logistic Regression, Random Forest, and Convolutional Neural Networks (CNN) with systematic optimisation through hyperparameter tuning to identify the best-performing model. The results indicate that CNN demonstrates the highest accuracy at 93.3% among the models studied. Furthermore, statistical tests confirm significant differences between the models, with CNN outperforming both LR-TFIDF and RF-BoW. The proposed model streamlines the exam preparation process and provides a standardised approach for educators and students, facilitating quicker insights into performance and more effective study strategies. This research contributes to the field by introducing an efficient and reliable automated classification system for exam questions, which can be integrated into various educational platforms, thereby improving educational practices by reducing the workload on educators and providing students with timely and accurate study materials.

TABLE OF CONTENTS

		Page
DECLARATION		iii
ACKNOWLEDGEMENT		iv
ABSTRAK		v
ABSTRACT		vi
TABLE OF CONTENTS		**vii**
LIST OF TABLES		x
LIST OF ILLUSTRATIONS		xi
LIST OF ABBREVIATIONS		xii
CHAPTER I	**INTRODUCTION**	
1.1	Introduction	1
1.2	Research Background	3
1.3	Problem Statement	4
1.4	Research Questions	5
1.5	Research Hypotheses	6
1.6	Research Objectives	6
1.7	Research Scope	6
1.8	Significance of the Study	6
1.9	Research Methodology	7
1.10	Thesis Organization	8
1.11	Conclusion	9
CHAPTER II	**LITERATURE REVIEW**	
2.1	Introduction	11
2.2	Machine Learning	11
2.3	Text Classification	12
2.4	Questions Classification using Machine Learning	14
	2.4.1 Logistic regression	27
	2.4.2 Random forest	27
	2.4.3 Convolutional neural networks	29
2.5	Feature Extraction Techniques	30

2.5.1 N-grams 30
2.5.2 Bag of words 30
2.5.3 Tf-idf 31
2.5.4 Word2vec 32

2.6 Conclusion 33

CHAPTER III METHODOLOGY

3.1 Introduction 34

3.2 Overview of Research Methods 34

3.3 Experiment Setup 36

3.4 Data Understanding 36

3.4.1 Questions from worksheet 36
3.4.2 Questions from past exam papers 37

3.5 Data Preprocessing 38

3.5.1 Dropping duplicate values 39
3.5.2 Case folding 39
3.5.3 Removal of unwanted symbols 39
3.5.4 Tokenisation 39
3.5.5 Stopwords removal 40
3.5.6 Removal of non-words 40
3.5.7 Lemmatisation 40

3.6 Dataset Description 40

3.7 Feature Extraction 41

3.7.1 Bag of words 41
3.7.2 Tf-Idf 42
3.7.3 N-grams 42
3.7.4 Word2vec 43

3.8 Machine Learning Development 43

3.8.1 Logistic regression 44
3.8.2 Random forest 44
3.8.3 Convolutional neural network 45

3.9 Data Partitioning (Cross-Validation) 46

3.10 Hyperparameters Tuning 47

3.10.1 Logistic regression 47
3.10.2 Random forest 48
3.10.3 Convolutional neural network 49

3.11 Evaluation 51

3.11.1 Performance metrics 51

3.12 Conclusion 52

CHAPTER IV RESULTS AND DISCUSSION

4.1 Introduction 53

4.2 Models Performance 53

4.2.1 Feature extraction technique 53

4.2.2 Optimisation of machine learning models through hyperparameter tuning 56

4.2.3 The best machine learning model for question classification 60

4.3 Conclusion 62

CHAPTER V CONCLUSION

5.1 Introduction 63

5.2 Summary 63

5.3 Implications 64

5.4 Limitations 64

5.5 Recommendation 65

5.6 Conclusion 66

REFERENCES 68

LIST OF TABLES

Table No. **Page**

Table 2.1 Literature Review Summary on Classifying Questions into Predefined Chapters 22

Table 3.1 Top Features using Bag-of-Words 42

Table 3.2 Top Features using TF-IDF 42

Table 3.3 Number of Features for each N-gram 43

Table 3.4 Values of Hyperparameter Tuning for Logistic Regression 48

Table 3.5 Values of Hyperparameter Tuning for Random Forest 49

Table 3.6 Values of Hyperparameter Tuning for CNN 50

Table 4.1 Accuracy of Different N-grams with LR and RF 54

Table 4.2 Logistic Regression Performance with Different Feature Extraction Methods 55

Table 4.3 Random Forest Performance with Different Feature Extraction Methods 55

Table 4.4 Performance of Convolutional Neural Network 55

Table 4.5 Hyperparameters Optimization Accuracy Performance 56

Table 4.6 Logistic Regression Best Model Parameters 57

Table 4.7 Random Forest Best Model Parameters 58

Table 4.8 Convolutional Neural Network Best Model Parameters 58

Table 4.9 Performance of the Best Models Before and After Hyperparameter Tuning 59

Table 4.10 Results for Classification Models Before and After Hyperparameter Tuning 59

Table 4.11 Average Processing Times Before and After Hyperparameter Tuning 60

Table 4.12 T-test Results After Comparing the Best Models 61

LIST OF ILLUSTRATIONS

Figure No. **Page**

Figure 1.1 Summary of CRISP-DM Tasks for this Research 8

Figure 2.1 Random Forest Model Schema 28

Figure 2.2 General Schema for CNN 29

Figure 3.1 Overview of Research Methods for this Study 35

Figure 3.2 Class Distribution of the Dataset 41

Figure 3.3 Structure of Random Forest for this Study 45

Figure 3.4 Structure of CNN Employed in this Study 46

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AQC	Automation Question Classification
BiLSTM	Bidirectional Long Short-Term Memory Networks
BoW	Bag-of-Words
BT	Bloom’s Taxonomy
CNN	Convolutional Neural Networks
CRISP-DM	Cross Industry Standard for Data Mining
GRU	Gated Recurrent Unit
IGCSE	International General Certificate of Secondary Education
KNN	K-Nearest Neighbours
LDA	Latent Dirichlet Allocation
LR	Logistic Regression
LSTM	Long Short-Term Memory Networks
LMS	Learning Management System
MFQC	Medical Forum Question Classification
MLP	Multi-Layer Perceptron
ML	Machine Learning
NB	Naïve Bayes
NLTK	Natural Language Toolkit Dataset
NLP	Natural Language Processing
QRNN	Quasi-Recurrent Neural Networks
QC	Question Classification
QAS	Question Answering System
QCS	Question Classification System
QP	Question Preprocessing
RF	Random Forest
RNN	Recurrent Neural Networks

STW	Supervised Term Weighting
SVM	Support Vector Machines
SwDA	Switchboard Dialogue Act Corpus
TC	Text Classification
USTW	Unsupervised Term Weighting
VB	Verb Extraction

CHAPTER I

INTRODUCTION

1.1 INTRODUCTION

Nowadays, an examination is the universal way of assessing students' performance during their scholarly pursuitx. Through exams, one can learn about their strengths and weaknesses. More importantly, it certifies whether students can proceed to the next stage of their academic journey (Guterman 2021). Recently, the exponential increase in the volume of past exam papers has posed significant challenges for educators and students in efficiently categorising and accessing relevant study materials. Manual classification of these questions into specific chapters is labour-intensive, time-consuming, and often inconsistent (Ko et al. 2012). This resulted in a growing interest in leveraging machine learning technologies to automate classification. Machine learning models have shown great promise in accurately categorising text-based data through the use of Natural Language Processing (NLP). Previous studies, such as those by (Aninditya et al. 2019; Baharuddin & Naufal 2023; Goh et al. 2023), have demonstrated the effectiveness of machine learning algorithms in classifying questions into pre-defined levels of cognitive abilities, thereby enhancing the efficiency and reliability of the classification process.

While various ways are used to gauge learning development, assessments through written exams are the most used method to evaluate the students' achievements (Mohammed & Omar 2020). Introduced more than 30 years ago, the International General Certificate of Secondary Education (IGCSE) is a popular international qualification for pupils aged fourteen to sixteen. It is favourably used in over 160 countries (Taqiyuddin & Aisyah 2023) and covers more than 70 subjects. IGCSE has three examination terms during the year: the first in February/March, then in May/June, and finally in October/November (Cambridge 2024). This results in an enormous number of exam papers for each subject since each subject may contain different exam papers targeting a specific question format. Moreover, each exam paper is released in multiple variants for each term, resulting in a high volume of past exams. While this benefits the students, it raises an unexpected problem: obtaining relevant practice questions is overwhelming for educators and students.

Traditionally, teachers played the role of bridging the gap by providing exam look-alike questions themselves. However, more is needed to prepare the students for the exams. Another method is for teachers to review the tremendous volume of question papers and design worksheets to aid the students. However, this takes much effort and time and is generally very hectic for teachers. Similarly, students need help searching available resources and identifying suitable questions aligned with specific topics and difficulty levels. This prompted the introduction of platforms that provide customised worksheets of practice questions. The problem with these platforms is that they manually categorise the questions. Manual classification may take a long time until it is updated on the platform or may not be accurate. Furthermore, some platforms may classify the same questions differently since different teachers performed the job. More importantly, due to multiple examination terms, manual classification of questions will be performed constantly throughout the year.

This research focuses on the critical gap between need and availability. This research proposes a model that accurately classifies questions according to pre-defined subject chapters'. Using powerful machine learning technologies will revolutionise exam preparation for IGCSE students and create a standard for classified questions. This model can later be integrated into a platform where an exam paper is uploaded, and it automatically classifies the questions and stores them in a database. Students can then search for their desired chapter and receive a worksheet with a specified number of questions. Teachers can easily create customised assessments, freeing them to focus on providing guidance and support to the students. Students can then gain quicker insights into their performance, allowing them to refine their study plans and strategies when approaching the exam.

This chapter will discuss the research background of the study related to exam preparations and questions classification, problem statement, research questions, research hypotheses, research objectives, research scope, the significance of the study, research methodology, and thesis organisation.

1.2 RESEARCH BACKGROUND

With advancements in technology, machine learning has significantly improved exam preparations. One of which is the introduction of adaptive learning platforms. These platforms provide students with different tools enabling personalised learning suiting their needs. A Learning Management System (LMS) was used by (Kim et al. 2023) to recommend questions to students based on questions they solved previously. The researchers further improved the systems by deploying a machine learning model to classify math questions accurately according to the difficulty level. With such an addition, students will receive questions of increasing difficulties rather than inconsistent complexity levels.

Developments in Natural Language Processing paved the way for Question Answering Systems (QAS). Machine learning-powered chatbots provide 24/7 on-demand assistance to students. These platforms have become integral in our lives, from answering simple questions to explaining complex concepts. While the number of internet users is increasing rapidly, more QAS are being developed for specialised usage. In (Mutabazi et al. 2023), a Medical Forum Question Classification (MFQC) system was proposed using improved deep-learning-based models. The questions were classified according to categories, which helped predict the answers to the medical questions while building the QAS.

Recently, Automation Question Classification (AQC) has been intensely researched. Humans can easily distinguish different questions and understand their meanings. However, it is more challenging for machines. Previous studies such as (Aburass & Dorgham 2023; Gani et al. 2022; Goh et al. 2023) mainly used Question Classification Systems (QCS) to categorise questions according to Bloom's Taxonomy (BT). BT is a model that organises educational goals based on their difficulty and specificity. Educators use the taxonomy to design learning activities and questions that cater to students at different levels.

Chapter classification of questions would help both educators and students of different needs. New questions are being generated daily worldwide, and with the tremendous amount of such resources, the classification will save educators and students some of their valuable time. Furthermore, the model could be integrated into LMS, QAS, and QCS alongside BT, further improving these existing systems.

1.3 PROBLEM STATEMENT

The manual classification of exam questions into pre-defined chapters is inefficient and lacks standardisation, posing challenges for students and educators preparing for the IGCSE exam. The current method needs to improve in terms of accuracy and consistency due to the diverse phrasing and complexity of questions. Existing studies concentrated on categorising questions according to Bloom's Taxonomy or custom taxonomies (Dachapally & Ramanam 2017; Momtazi 2018; Pota et al. 2020), but there is a notable absence of research targeting classification based on subject-specific-chapters classification. This research gap calls for the proposal of a robust model that can accurately classify exam questions into pre-defined chapters, streamlining the exam preparation process and providing standardised resources for students and educators. This can potentially aid students' learning experience and academic performance.

Effective feature extraction techniques are essential for enhancing the accuracy of question classification tasks in natural language processing. While previous studies have explored multiple feature extraction methods such as bag-of-words (Momtazi 2018), term weighting schemes (Gani et al. 2022), and word embedding models like word2vec (Luo et al. 2021; Zulqarnain et al. 2021), their suitability for classifying exam questions according to subject's chapters remains unexplored. This research aims to identify and propose the most compelling feature extraction techniques explicitly tailored for enhancing question classification accuracy, addressing the need for reliable and efficient classification models in exam preparation.

Hyperparameter tuning is crucial in optimising the machine learning models' performance, including those used for question classification tasks. Research such as (Mutabazi et al. 2023; Pota et al. 2020; Seidakhmetov 2020; Zulqarnain et al. 2021) fine-tuned deep learning models through hyperparameters optimisations. However, existing studies in this domain have yet to explicitly address using hyperparameter tuning to enhance machine learning or ensemble models in question classification tasks. This study seeks to address this omission by optimising the performance of the proposed classification model through systematic hyperparameter tuning. This study aims to provide practical insights into effectively utilising hyperparameter tuning techniques in machine learning-based question classification systems by maximising model performance, improving generalisation, and enhancing predictive accuracy.

In summary, while previous research has primarily focused on taxonomies like Bloom's Taxonomy, there is a notable gap in addressing classification based on subject-specific chapters. This research aims to propose a robust model for accurately classifying exam questions into chapters. Furthermore, this study seeks to explore the performance of some feature extraction methods to address this gap. Finally, explicit utilisation of hyperparameters tuning for machine learning and ensemble models in enhancing classification models for exam questions has been overlooked. By systematically optimising hyperparameters, this research aims to maximise model performance, improve generalisation, and provide practical insights into enhancing machine learning-based question classification systems.

1.4 RESEARCH QUESTIONS

The research questions of the study are derived from the problem statements as follows:

1. Which machine learning model most effectively classifies questions according to chapters?
2. Which feature extraction is most suited for question classification according to chapters?
3. How does hyperparameter tuning impact the performance of the models?

1.5 RESEARCH HYPOTHESES

Based on the problem statements, the study proposes the following research hypotheses:

1. Convolutional Neural Networks will outperform the other models with their ability to capture the context and dependencies within text.
2. Term Frequency-Inverse Document Frequency (TF-IDF) highlighting the importance of words in the question will produce significantly better results than other feature extraction techniques.
3. The developed classification models' performance will be improved through hyperparameter tuning.

1.6 RESEARCH OBJECTIVES

This study aims to achieve the following objectives:

1. To propose the best model that accurately classifies questions into pre-defined chapters.
2. To propose the most effective feature extraction method for enhancing question classification accuracy.
3. To optimise the performance of the machine learning model through hyperparameter tuning.

1.7 RESEARCH SCOPE

This research focuses on Physics (0625) from the IGCSE curriculum. This research will not explore subjective or open-ended questions; instead, it will target multiple-choice questions only. The scope includes a thorough evaluation of the performance in classifying previous exam questions.

1.8 SIGNIFICANCE OF THE STUDY

This research importance is its ability to address critical issues in the realm of exam preparation for students; it includes:

1. Revolutionise exam preparation by automatically leveraging machine learning to classify past exam questions into subject chapters. This offers a solution to the problem faced by various educators and students in obtaining relevant practice questions. The model will also ease labelling new questions when new exam papers are released throughout the year.
2. Standardising a system for classifying questions will reduce the variances in categorisation resulting from multiple teachers manually labelling past exam questions.
3. The model could adapt to different international and domestic qualifications.
4. Using a machine learning-based system for categorising questions set the stage for similar applications in other domains.

1.9 RESEARCH METHODOLOGY

The research design of this study follows the Cross Industry Standard for Data Mining (CRISP-DM) framework. CRISP-DM consists of six sequential phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. The overview of the tasks in each phase is shown in Figure 1.1, and the outline of each phase in this study is as follows:

1. Business understanding: In this phase, existing research gaps in the domain of question classification are identified. It involves defining the problems to be solved, determining the research objectives, and converting these objectifies into data mining problem definitions. More research needs to be conducted targeting classification based on subject-specific chapters; this research aims to propose a machine-learning model capable of accurately classifying questions. Furthermore, this study seeks to identify the best feature extraction method for the task and optimise the model's performance through hyperparameter optimisation.
2. Data understanding: In this phase, exam papers are collected, and questions are extracted and explored to uncover insights and identify potential data quality issues.
3. Data preparation: In this phase, the final dataset used for modelling is constructed from the raw dataset. This involves correcting issues identified in the previous step

and using various NLP techniques to prepare the data for machine-learning model development.

4. Modelling: Machine learning techniques are selected and applied to the prepared dataset during this phase. This phase also involves enhancing the model's performance through hyperparameter optimisation.

5. Evaluation: The models are evaluated in this phase to ensure they meet the goals set during business understanding. This includes verifying that the models provide a reliable and consistent classification of questions into chapters.

6. Deployment: Generally, this phase involves deploying the model into an operational environment. However, this phase of this research consists of presenting and comparing the results from the models.

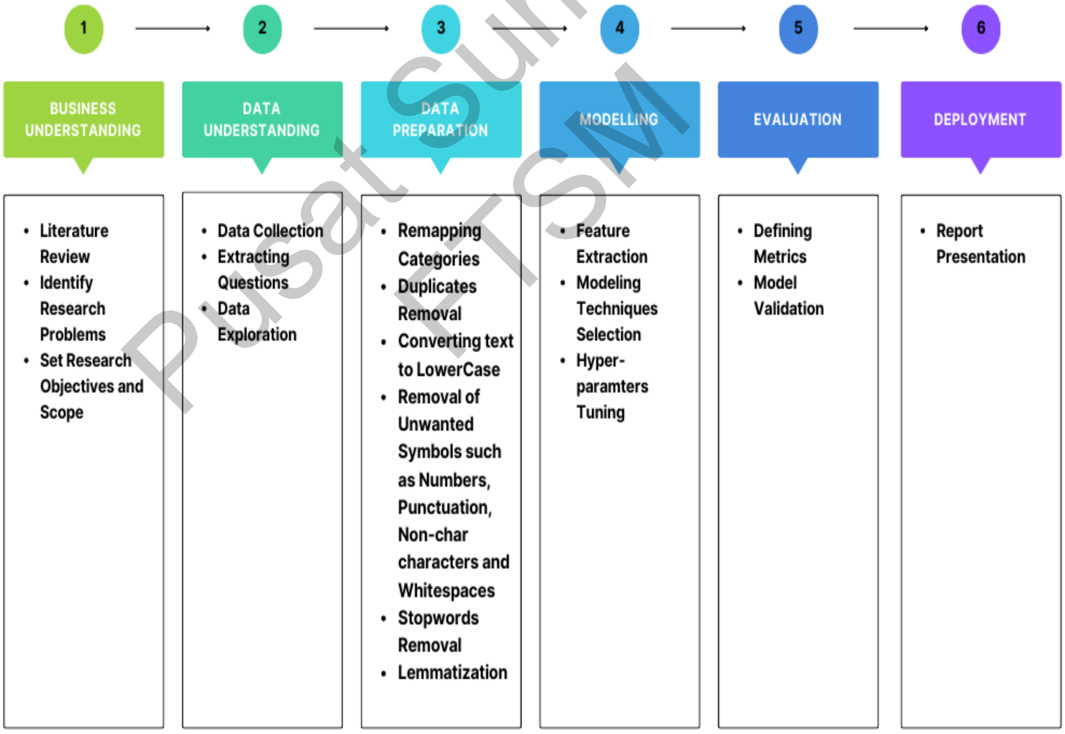

Figure 1.1 Summary of CRISP-DM Tasks for this Research

1.10 THESIS ORGANIZATION

This thesis consists of five chapters structured as follows:

Chapter 1 introduces teachers' and students' difficulties when preparing for the IGCSE exam. The chapter also discusses the issues with the current solutions aiding exam preparation. Additionally, this chapter covers the research background, problem statements, research questions, research hypotheses, research objectives, research scope, significance of the study, and methodology.

Chapter 2 critically reviews current literature on question classification using machine learning models. By examining the literature and identifying gaps in knowledge, this chapter presents the theoretical framework that guides this thesis.

Chapter 3 centres on the research methodology employed in the study, outlining the research design, data collection methods, pre-processing techniques, feature extraction methods, and the machine learning models developed. Furthermore, data analysis techniques and evaluation metrics for assessing the performance of the models will be discussed.

Chapter 4 presents the study's findings, highlighting the best machine learning model for question classification according to subject-specific chapters and the best feature extraction technique. Moreover, the chapter shows how hyperparameter tuning affects the models' performances.

Chapter 5 summarises the study, addresses its limitations of the study, and suggests possibilities for future research.

1.11 CONCLUSION

This chapter discussed one of the challenges faced by students preparing for the IGCSE exam and the limitations of the current solution, which laid the foundation for the research background. Exam questions will be generated continuously as long as there is a written assessment. With such an extensive database of questions, classifying questions based on the subject's chapters will aid both students and educators. Furthermore, this classification could be implemented alongside existing systems, such as LMS, to improve students' overall performance. Previous studies have used different techniques for classification; therefore, this research aims to test a machine learning model, an ensemble model, and a deep learning model for classifying questions into pre-defined chapters.

Additionally, this study aims to evaluate multiple feature extraction techniques and improve the model's performance using hyperparameter tuning. The study’s scope is objective questions in physics, a subject of the IGCSE curriculum. This study will extract the features using N-Grams, Bag-of-Words, TF-IDF, and Word2vec. Furthermore, the models to be developed are Logistic Regression, Random Forest, and Convolutional Neural Networks.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

In this chapter, a thorough examination of existing literature regarding question classification into pre-defined subject chapters using machine learning techniques is performed, exploring theoretical frameworks, recent research, and areas needing further exploration. This literature review includes a variety of research conducted in numerous countries and multiple applications of question classification.

2.2 MACHINE LEARNING

Machine learning (ML) has witnessed substantial growth and innovation in recent years, reshaping various industries and technological landscapes. Machine Learning is a branch of Artificial Intelligence (AI) aiming to develop computer models capable of making decisions autonomously (Kufel et al. 2023). These models continuously enhance their accuracy by learning from data; therefore, the dataset's quality affects the model's performance enormously. Two fundamental paradigms of machine learning are supervised and unsupervised learning.

During supervised machine learning, the model is trained on a labelled dataset, with each input data point matched to its corresponding output data. The model learns the relationship between the two entities and how to map inputs to outputs. The model can then make predictions and decisions using unseen data (Dake et al. 2023; Kim et al. 2023). This type of learning is categorised into two applications: prediction and classification. In classification, the model maps inputs to a certain number of class labels. Examples of classic supervised learning classification include email spam detection or image classification tasks. In regression problems, the model predicts continuous values using input data. In finance, using historical data and market indicators, a regression model could be utilised to predict stock prices (Kufel et al. 2023). In unsupervised ML, the model uses unannotated data. This algorithm groups the data based on its features; this is achieved by uncovering hidden patterns or structures within the data (Aninditya et al. 2019). Unsupervised learning is categorised into clustering and association.

The need for machine learning in classifying topics on past year exams arises from the growing volume of data and the necessity for efficient analysis (Zulqarnain et al. 2021). Traditional methods of categorising exam questions are time-consuming and prone to human error. Previous studies have demonstrated the efficacy of machine learning techniques in automating this process, significantly reducing the workload for educators and pupils (Su et al. 2022). Researchers have employed various machine learning algorithms to categorise exam questions, showcasing the potential for these techniques to improve accuracy and consistency. Machine learning techniques like Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and Long Short-Term Memory Networks (LSTM) have been extensively studied for question classification tasks, achieving high accuracies of up to 93.7% (Mohasseb et al. 2018; Pota et al. 2020; Seidakhmetov 2020). Natural language processing and multi-label classification approaches have been successfully used to categorise complex textual data into relevant topics, highlighting the importance and efficiency of machine learning in educational contexts (Goh et al. 2023).

This study focuses on developing machine learning models to classify past exam questions automatically into pre-defined subject-specific chapters. To achieve this, the study concentrates on multi-label classification within text classification. Machine learning-based text classification will be discussed in section 2.3.

2.3 TEXT CLASSIFICATION

Text classification (TC), or document classification, assigns pre-defined labels to text or a corpus based on its content (Aninditya et al. 2019). This procedure could be accomplished manually, but it is labour-intensive and time-consuming. Moreover, manual classification is prone to errors due to human misinterpretation or lack of domain knowledge. In today’s era of big data, an enormous amount of digital data is available daily, making it challenging for TC. There was a vast revolution when machine learning replaced manual work because it saves time and is highly accurate in classification tasks. Text classification has been applied in various fields, including internet page classification, author attribution, management of knowledge, and detecting spam. With recent developments in NLP, TC has been used in chatbots, sentiment analysis, service recommendation, search optimisation, and more (Palanivinayagam et al. 2023; Zhu & Lei 2022).

Text classification using machine learning comprises four key stages: (a) preprocessing, (b) text representation, (c) feature selection, and finally (d) classification (Palanivinayagam et al. 2023). Text preprocessing involves the removal of noise from the input data. Text is usually preprocessed with tokenisation, lemmatisation, or stems. Depending on the nature of the task, preprocessing may include stop word removal, punctuation removal, part-of-speech tagging, or transforming the text into lowercase in preparation for the next step. A classical text representation model is usually obtained through Bag-of-Words (BoW) or n-grams. Word embeddings and topic modelling have recently gained popularity in text representation. Feature selection is an optional step that reduces the number of features, thus reducing noise (Palanivinayagam et al. 2023). The most widely used methods include Term Frequency-Inverse Document Frequency, Chi-square Statistics, Information Gain, and Mutual Information (Zhu & Lei 2022), followed by modelling or classification tasks. Traditionally, popular classifiers have included Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision Trees, Random Forests, and Support Vector Machines (Mohammed & Omar 2020; Mohasseb et al., 2018). Recently, Classifiers based on deep learning have shown remarkable results by capturing complex non-linear relationships within the data, outperforming traditional machine learning techniques (Liu 2022; Seidakhmetov 2020).

Various question classification applications are discussed in section 2.4, and an extensive review of past studies on machine learning-based question classification will be conducted.

2.4 QUESTIONS CLASSIFICATION USING MACHINE LEARNING

Question classification (QC) using machine learning techniques is an essential task within NLP. The task aims to classify questions into predefined labels or categories automatically. Question answering systems are affected by the quality of question classification systems. A QAS is explicitly designed to offer accurate and relevant responses to user inquiries in natural language by obtaining information from a specific set of documents (Liu 2022). Through question classification, these systems offer information and criteria to guide the answer selection strategy. They categorise answer types and precisely identify and validate the answers (Su et al. 2022). QC is also highly used in Question Classification Systems. QCS, used in the education industry, mainly focuses on classifying questions based on BT. By aligning the course learning outcomes and BT, educators can evaluate students' performance and identify their weaknesses. Questions classification has also been integrated into Learning Management Systems, an online database that stores student information. Through incorporating QC, these systems can suggest suitable questions to students, leveraging their past responses and proficiency level for personalised learning experiences (Kim et al. 2023).

(Kim et al. 2023) classified mathematical questions into difficulty levels and determined the critical features for the task using mathematical test items provided by ABLE Tech. The researcher experimented with 13 machine learning models and determined Xgboost had the highest accuracy of 85.7%, surpassing the other models. The model's optimal parameters were obtained via a grid search with cross-validation. The study also concluded a positive correlation between the rate of correct answers, each section of the question, and the solution time. However, there was no correlation with the answer type of the question. This research presents a breakthrough in the question classification domain; however, it is only limited to one subject, making its scalability questionable. Another area for improvement lies in the variables pre-extracted within the selected dataset, which may be challenging to implement with other datasets.

(Pota et al. 2020) conducted numerous experiments to assess the influence of several factors and hyperparameters on CNN classification performance. The dataset used consisted of questions from both English and Italian derived from the TREC dataset, and the questions were classified according to a custom taxonomy by the researcher. For the English dataset, the model scored an accuracy of 93% after hyperparameters optimisation. In terms of text representation, it was found that avoiding punctuation will yield better results alongside using pre-trained word embedding vectors with dimensions equal to 300. The model's architecture should consist of 100 filters of size 2, and infinitely increasing functions or identity is preferred. Small batches of 10 shall be used for learning. Meanwhile, dropouts and loss regularisation should be evaded. Finally, Adadelta is the best optimiser, with a learning rate 1. CNNs are widely favoured for text classification because of their exceptional performance results. This study presents valuable insights into the best parameters when building the model.

(Mutabazi et al. 2023) they proposed an improved medical form question classification model using CNN and Bidirectional Long Short-Term Memory Networks (BiLSTM). The study's dataset was the ICHI Dataset and MedQuAD Dataset. The model combined the strengths of CNN and BiLSTM with CNN for feature extraction and BiLSTM for sequence classification. The first step was removing stop words and unwanted symbols, followed by word2vec embedding—and finally, the features using CNN and BiLSTM for sequence learning. The model surpasses all baseline models on both datasets used during the study, with 57% and 93% accuracy, respectively. However, relying on word2vec could lead to problems with out-of-vocabulary or domain words.

(Aburass & Dorgham 2023) they employed a Dual-branch neural network where each branch processes a different type of embedding: one uses Electra, and the other uses GloVe. LSTM layers were then applied to the combined embeddings to achieve the final categorisation of questions from the TREC dataset. Compared to the baseline models, the proposed method’s performance was much superior, with 80% accuracy. For pre-processing, the text was converted into lowercase. The input text underwent two tokenisation processes: one for Electra and another for GloVe. A fixed sequence length of 512 was also enforced through padding. Even though the model displayed impressive results compared to the baseline models used in the study, the model is still far from the performance obtained by using CNN (Pota et al. 2020).

(Zulqarnain et al. 2021) they investigated different deep learning techniques and combinations to determine which is most suited for classifying questions in Turkish, adapted from an English Question Dataset (UIUC). The researcher implemented three primary deep learning techniques: LSTM, Gated Recurrent Unit (GRU), and CNN. Another two hybrid models were tested, a combination of CNN and GRU (CNN-GRU) and a combination of CNN and LSTM (CNN-LSTM). For word embeddings, the Word2vec method was used, employing skip-gram and Continuous Bag-of-Words methods with different vector sizes. Results showed that the two combinational models performed better with skip-gram rather than CBOW. The CNN model, utilising skip-gram with 300 feature vectors, achieved the highest accuracy at 93.7%. The study concluded that Word2vec models successfully capture both semantic and syntactic relationships among words, thereby improving the effectiveness of classification models. Since the questions were general and not domain-specific, word2vec performed well.

(Goh et al. 2023) proposed a rule-based semantic method to classify 200 diploma course questions from the University of Wollongong Malaysia KDU University College according to Bloom’s Taxonomy. Lecturers manually categorised the questions into single-sentence and multi-sentence questions. The proposed system is constructed from three different modules, commencing with question preprocessing (QP); the system utilised tools such as Natural Language Toolkit (NLTK), Stanford POS tagger, and WordNet similarity approaches for processing. The following module is the verb extraction (VB), where verbs are identified from the input text and compared with Bloom’s Taxonomy verb list using a similarity to determine the question’s category. The researchers concluded that wordnet with wu-palmer semantic similarity outperformed other methods with 83% accuracy.

(Gani et al. 2022) performed a comparative analysis between supervised and unsupervised term weighting schemes to classify previous exam questions according to BT using three different datasets. For pre-processing, the questions were converted into lowercase, tokenised, punctuation removed, stop words removed, applied lemmatisation, and finally, part of speech tagged. The study experminted on three Supervised Term Weighting (STW) schemes and three Unsupervised Term Weighting (USTW) schemes. In terms of STW, both TF-IDF-ICF and TF-IDF-ICSDF were outperformed by TF-ICF. Moreover, TF-ICF was the best scheme used during the experiments. Regarding the USTW schemes, TF-IDF fell short behind both TFPOS. The study also suggested that integrating part-of-speech-based weighting and document distribution by class categories could enhance classification performance. Regarding the ML models used, The Multi-layer perceptron (MLP) classifier, when used with the TF-ICF term weighting scheme, consistently outperformed both Naïve Bayes (NB) and SVM across all datasets. SVM with TF-ICF also showed significant performance improvements compared to other term weighting schemes, especially in multi-domain datasets.

(Momtazi 2018) introduced a novel approach for categorising questions within community-based question-answering systems. The dataset utilised in the study contained 1000 questions in German and 2800 in Persian. The model utilised latent semantics from the text to classify questions by assessing the relationship between the topics and labels. This is accomplished through Latent Dirichlet Allocation (LDA). The proposed method surpassed the baseline algorithms in terms of accuracy. The study also highlights how the proposed method could be used for other categorical text classifications, indicating its generalizability.

(Dachapally & Ramanam 2017) they used a two-tier CNN model to improve efficiency and reduce the time spent on question classification. The first tier of CNN was used to predict the primary topic of the question, while the secondary tier predicted the sub-topic. The model was trained using 5452 questions categorised into 6 primary topics and 50 sub-topics from the University of Illinois. The model was tested on the TREC dataset and a dataset of 115 manually collected questions from the Quora website. The study concluded that the model trained solely on word2vec adapted more effectively to unseen examples than those trained with both word2vec and GloVe. The model recorded 90.4% Main category accuracy and 76.5% Sub-category accuracy. This may be attributed to the extensive vocabulary range in word2vec (1.2 million) compared to GloVe (400,000).

(Su et al. 2022) addressed the challenges of obtaining a considerable number of labelled data particularly in new domains. The study evaluated the performance of various machine learning models for cross-domain question classification. The Yahoo! Answers dataset, classified into 10 categories, was used in the study. Semantic information from the category labels was also obtained alongside WordNet to expand the question, further improving classification accuracy. L-ALBERT-FiT achieved 81.8% average accuracy in cross-domain question classification, increasing to 86.7% after dataset expansion. The study highlights that the model performs poorly, and experimental recall decreases when there is an uneven distribution of target domain categories. Additionally, introducing synonyms significantly enhances accuracy when the initial accuracy rate is low. However, once accuracy is high, the added noise from text expansion can decrease classification accuracy.

(Liu 2022) proposed a method for extracting multidimensional features to achieve accurate classification of medical questions. This is done by integrating multiple neural networks for feature extraction to improve classification results. The medical questions used in the experiment were collected from the 120ask question answering community. The dataset had 17,387 questions categorised into 20 categories. The proposed model was able to extract more features and obtain better results than simple Recurrent Neural Networks (RNN) and LSTM. However, the experiment displayed lower results than simple GRU, with 54% to 56.5% accuracy. This can be attributed to the poor performance of RNN (32.6%), which affects the final features.

(Seidakhmetov 2020) used the TREC dataset and compared the performance of Logistic Regression (LR), BI-LSTM, CNN, and Quasi-Recurrent Neural Networks (QRNN) for question classification tasks. Regarding pre-processing, the input text was tokenised and converted to word indices. Word embedding was performed using GloVe and with tokens mapped into a 300-dimensional vector. The Convolutional Neural Network-based approach, utilising kernel sizes of 2, 3, 4, 5, and 6 along with a single fully connected layer, achieved the best result with an accuracy of 90.7%. This further solidifies CNN’s remarkable performance in question classification. With hyperparameter optimisation, the model's performance should reach the results obtained by (Pota et al. 2020).

(Mohammed & Omar 2020) classified exam questions based on BT. The study utilised two datasets, both categorised into six classes: one contained 141 open-ended questions from multiple sources, while the other was sourced from a previous study and contained 600 questions. Several experiments were conducted using three classifiers and three features. The features included Term Frequency-Inverse Document Frequency (TF-IDF), Term Frequency with Part-of-Speech (TFPOS) and Inverse Document Frequency (TFPOS-IDF), and Word2Vec combined with TFPOS-IDF (W2V-TFPOSIDF). The classifiers used were K-Nearest Neighbors (KNN), LR and SVM. The combination of SVM and W2V-TFPOSIDF scored the highest F1-measure on both datasets (83.7% and 89.7%). The study also showed significant improvements between TF-IDF versus TFPOS-IDF and TFPOS-IDF versus W2V-TFPOSIDF, indicating the resilience of the proposed model.

(Mohasseb et al. 2018) a question classification framework was proposed that can adapt to various question-answering systems by developing domain-specific grammatical rules and patterns tailored to each question type. Five thousand randomly selected questions from a collection of datasets were used for the study. Decision Tree outperformed the other models with 86% accuracy.

(Aninditya et al. 2019) they classified exam questions between 2012 and 2019 according to BT using Naïve Bayes as a classifier. Tokenisation, stemming, and filtering were used to pre-process the text. The researcher experimented with 3 feature extraction techniques. Using Naïve Bayes with features TF-IDF and n-grams improved the model's accuracy to 85% precision and 80% recall.

(Luo et al. 2021) proposed a model that involves a self-attention algorithm for encoding the question and a combination of RNN to classify the category of the question. The model was trained and tested on The Switchboard Dialogue Act Corpus (SwDA) and the NLTK dataset. The proposed model was compared with various models from other studies and outperformed them all on both datasets except once, with 83.1% and 85.5% accuracy.

(Ponce-L´opeza 2024) evaluated the utility of LLMs in classifying multi-choice questions related to medical subjects. The MedMCQA dataset was used for the study. The dataset contained 183k questions for training, 6k for development, and 4k for testing. The researcher used Sentence-BERT for multi-sequence classification to extract sequence embedding and fine-tune them for MCQA tasks. The proposed model acquired a 60% test accuracy.

(Aithal et al. 2021) they proposed an application that combines question generation with a question-answering system. ProphetNet generated question-answer pairs, while the answers were generated using BERT. Each query from the SquAD 2.0 dataset resulted in multiple question-answer pairs. A question similarity mechanism classified each question as answerable, unanswerable, or irrelevant based on cosine similarity with the original query. This approach enhances the efficiency of question-answering systems by focusing exclusively on pertinent and answerable questions. The model had an efficiency of 48% of unanswerable questions, meaning the model could not answer these questions on the Squad 2.0 dataset and 91% on the SquAD 1.1 dataset. However, the model had 100% efficiency on all irrelevant questions because the model could not identify irrelevant questions.

(Bae & Ko 2019) proposed a novel model for retrieving questions based on generating the case frame of a sentence, utilising word embedding to calculate similarities between sentences, and addressing the lexical gap problem with word embedding. The research did not explicitly mention the ML model used for question categorisation and retrieval. The researcher compared novel weighting methods with TFIDF and determined that the model is adequate for question retrieval.

(Baharuddin & Naufal 2023) they explored the capabilities of IndoBert in classifying Indonesian multiple-choice questions based on BT. 449 MCQ questions from the elementary school level from previous sources were used as the dataset. The model showed impressive results with 97% accuracy.

Table 2.1 presents a summary of the literature review on the application of machine learning for classifying questions into predefined labels. According to previous studies, deep learning models consistently outperformed other techniques in terms of performance. Most notably, convolutional neural networks and their variations were the best-performing models in accuracy. Various ensemble learning techniques were tested, where Xgboost and Random Forrest excelled. Meanwhile, some studies have used Logistic regression as a baseline model. Therefore, in line with these findings, Convolutional Neural Networks, Random Forest, and Logistic Regression are implemented in this thesis to classify questions into pre-defined chapters

Table 2.1 Literature Review Summary on Classifying Questions into Predefined Chapters

No.	Author	Objective	Data	Algorithm	Result
1	(Kim et al. 2023)	Classify the difficulty of mathematical questions and determine the most relevant features that has significant effect on questions difficulty.	Mathematical test set from ABLE Tech	LR KNN SGD CART SVM MLP RF ET ADA CatBoost GBM LightGBM XGB	With an accuracy of 85.7%, XgBoost surpasses the other models. In terms of features, the study concludes that there is no correlation between the answer type. Furthermore, there is a positive correlation with the solution time, the rate of correct answers and each section of the items.
2	(Pota et al. 2020)	Analyze and optimise Convolutional Neural Networks settings for question classification.	A dataset of both English and Italian questions derived from the TREC dataset	CNN	After optimizing the hyperparameters, the model had an accuracy of 93% on the English dataset. In terms of text representation, it was found that avoiding punctuation will yield better results alongside using a pre-trained word embedding vectors with dimension equals to 300. The architecture of the model should consist of 100 filters of size 2, also the use of infinitely increasing functions is preferred or identity. Small batches of 10 shall be used for learning, meanwhile dropouts and loss regularization should be avoided.
3	(Mutabazi et al. 2023)	Propose an improved medical form question classification model.	ICHI Dataset and MedQuAD Dataset	CNN BiLSTM Proposed model	The proposed model outscores the baseline models on both datasets used during the study with an accuracy 57% and 93% respectively.

to be continued …

… continuation

4	(Aburass & Dorgham 2023)	Improve the performance of question classification by combining the capabilities of Electra, GloVe and LSTM.	TREC dataset	Electra BERT RoBERTa Distilbert Proposed Model	Compared to the baseline models, the proposed method performance was much more superior with 80% accuracy.
5	(Zulqarnain et al. 2021)	Investigating deep learning methods to classify questions in the Turkish language.	Turkish questions that are adapted from an English Question Dataset (UIUC)	GRU LSTM CNN CNN-GRU CNN-LSTM	The CNN model, utilizing skip-gram with 300 feature vectors, achieved the highest accuracy at 93.7%. The study concluded that Word2vec models successfully capture both semantic and syntactic relationships among words, thereby improving the effectiveness of classification models.
6	(Goh et al. 2023)	A question classification system that uses a semantic and synthetic approach to accurately classify examination questions based on BT.	200 Engineering diploma course questions (Single and multiple Sentence types) from UOW Malaysia KDU University College	Rule-based semantic approach	Wordnet with wu-palmer semantic similarity outperformes other measures with 83% accuracy.
7	(Gani et al. 2022)	Conduct a comparative analysis between supervised term weighting and unsupervised term weighting schemes for exam question classification based on BT.	181 business questions and 415 questions from various field and 600 questions from a previous research	NB SVM MLP	In terms of STW, both TF-IDF-ICF and TF-IDF-ICSDF were outperformed by TF-ICF. Moreover, TF-ICF was the best scheme used during the experiments. Regarding the USTW schemes, TF-IDF fell short behind both TFPOS. The study also suggested that integrating POS-based weighting with document distribution according to class categories might enhance the performance of exam question classification.
8	(Momtazi 2018)	Introduce a novel appproach to classify community-based question answering systems questions.	1000 questions in German and 2800 questions in Persian	SVM NB LDA	The proposed method surpasses the baseline algorithms in terms of accuracy. The study also highlights how the proposed method could be used for other categorical text classifications, indicating its generalizability.

to be continued …

… continuation

9	(Dachapally & Ramanam 2017)	Simplify question classification using a two-tier CNN.	For training 5452 questions from University of Illinois. For testing TREC dataset and manually collected 115 questions from Quora website	CNN	The study concluded that the model trained solely on word2vec adapted more effectively to unseen examples compared to the model trained with both word2vec and GloVe. The model recorded 90.4% Main category Accuracy and 76.5% Sub-category Accuracy.
10	(Su et al. 2022)	Improve question classification accuracy in new domains through deep transfer learning methods.	The Yahoo! Answers dataset categorized into 10 categories	SVM KNN NB ULMFIT Flair XLNet Bert ALBERT Bert-fit Albert-fit L-Albert-fit	L-ALBERT-FiT achieved 81.8% average accuracy in cross-domain question classification, increasing to 86.7% after dataset expansion.
11	(Liu 2022)	Propose a multi-dimensional feature extraction-based model for question classification.	The 120ask question answering community. Contains 17,387 question data in 20 categories	RNN LSTM GRU Proposed model	The proposed model was able to extract more features and obtained better results compared to simple RNN and LSTM. However, the experiment displayed lower result compared to simple GRU with 54% to 56.5% accuracy respectively. This can be attributed to the poor performance of RNN (32.6%) which affects the final features.
12	(Seidakhmetov 2020)	Provide a comparative study multiple approaches in question classification task.	TREC dataset	LR Bi-LSTM CNN QRNN	CNN-based approach, utilizing kernel sizes of 2, 3, 4, 5, and 6 along with a single fully connected layer, scored the highest result with an accuracy of 90.7%.

to be continued …

… continuation

13	(Mohammed & Omar 2020)	Classify exam questions based on BT cognitive domain, specifically focusing on multiple domains rather than just a single domain.	Two open domain datasets. 141 and the other is 600 open-ended questions	KNN LR SVM	Combination of SVM and W2V-TFPOSIDF scored the best F1-measure on both datasets (83.7% and 89.7%). The study also showed a significant improvements between TF-IDF versus TFPOS-IDF and TFPOS-IDF versus W2V-TFPOSIDF, indicating the resilienceof the proposed model.
14	(Mohasseb et al. 2018)	A question classification framework that can adapt to various question-answering systems.	5,000 questions were randomly selected from a collection of datasets	SVM RF DT NB	Decision Tree outperformed Random Forrest, Support Vector Machines and Naive Bayes with 86% accuracy.
15	(Aninditya et al. 2019)	Classify exam questions based on BT.	Exam questions between 2012 - 2019	NB	Using Naive Bayes with features TF-IDF and n-grams improved the accuracy of the model to reach 85% precision and 80% recall.
16	(Luo et al. 2021)	Improve question and answer problem classifications by utilizing a Deep Contextualized Transformer model.	Switchboar d Dialogue Act Corpus (SwDA) and The Natural Language Toolkit Dataset (NLTK)	Proposed model	The proposed model outperformed the other models on both datasets except once with 83.1% and 85.5% accuracy.
17	(Ponce-L´opeza 2024)	Evaluate the utility of LLMs in classifying multi-choice questions related to medical subjects.	MedMCQA dataset	BERT BioBERT SciBERT PubmedBERT Codex Proposed model	The proposed model acquired a 60% test accuracy.
18	(Aithal et al. 2021)	An application which combines Question Generation and question Answering Systems.	SQuAD 2.0 dataset	BERT	The model had an efficiency of 48% of unanswerable questions on the Squad 2.0 dataset and 91% in SquAD 1.1 dataset. However, the model had a 100% efficiency on all irrelevant questions because the model couldn’t identify irrelevant questions.

to be continued …

… continuation

19	(Bae & Ko 2019)	Improve question classification and retrieval performance on cQA services.	4,702 question- answer pairs in Naver KIN, a cQA in Korea.	The paper does not explicitly mention the specific machine learning model used for question classification and retrieval.	The researcher compared a novel weighting methods with TFIDF and determined its effectiveness in question retrieval.
20	(Baharuddin & Naufal 2023)	Build classification system to classify Indonesian exam questions according to BT	449 extracted MCQ question for Indonesian elementary school level exam	IndoBERT	The model showed impressive results with 97% accuracy.

2.4.1 Logistic Regression

Logistic Regression is a statistical method illustrating the relationship between a dependent variable and one or more independent variables. It uses logistic or sigmoid functions to predict one of two possible outcomes. This approach transforms the linear combination of independent variables into a probability score between 0 and 1 (Kufel et al. 2023). Logistic Regression is widely used in diverse machine learning applications, including text classification, and has proven to be highly effective. For multiclass classification, Logistic Regression uses the one-vs-all strategy, applying maximum likelihood estimation to determine the predicted class:

$$\boldsymbol{p(y = 1|x) = 1 \div (1 + e^{-z})} \qquad ..(2.1)$$

where p(y = 1|x) represents the probability of the binary outcome (y = 1) given the independent variables (x), and z = b0 + b1x1 + b2x2 + . . . + bnxn represents the linear combination of the independent variables (x) and their coefficients (b), where b0 is the intercept and b1 to bn are the coefficients of the independent variables x1 to xn. The function e is the base of the natural logarithm. The logistic function on the right-hand side of the equation ($1/(1 + e^{-z})$) maps the linear combination of the independent variables (z) to a probability value between 0 and 1, which represents the predicted probability of the binary outcome (y = 1) given the independent variables (x) (Kim et al. 2023). The probability is given by:

$$\boldsymbol{p = \frac{1}{1 + e^{-(\mathrm{b0\ +\ b1x1\ +\ b2x2\ +\ \ldots +\ bnxn})}}} \qquad ..(2.2)$$

2.4.2 Random forest

Random Forest (RF) is a flexible ensemble technique applied in machine learning for both classification and regression tasks. It integrates multiple independent decision trees during the training phase. Each decision tree divides the data into subsets based on feature values and makes predictions at each node to identify the target variable. The Random Forest's output is generated by aggregating the predictions from these individual trees: averaging the predictions for regression tasks and using a majority vote for classification tasks. RF is highly effective for large datasets with high dimensionality, recognised for its accuracy and resilience to noise and overfitting. Mathematically, the output of Random Forest for a new sample x is defined as follows:

$$\boldsymbol{RF(x)} = \begin{cases} \boldsymbol{\frac{1}{T}\sum_{t=1}^{T} f(x)} & \boldsymbol{Regression} \\ \boldsymbol{mode\big(f_1(x), f_2(x), \dots, f_T(x)\big)} & \boldsymbol{Classification} \end{cases} \quad ..(2.3)$$

Where T is the number of decision trees, and $f_T(x)$ is the prediction of the *t*-th decision tree for the input x. The randomness introduced in Random Forest helps reduce over-fitting and improve the model's accuracy (Kim et al. 2023). The framework of a random forest model is shown in Figure 2.1.

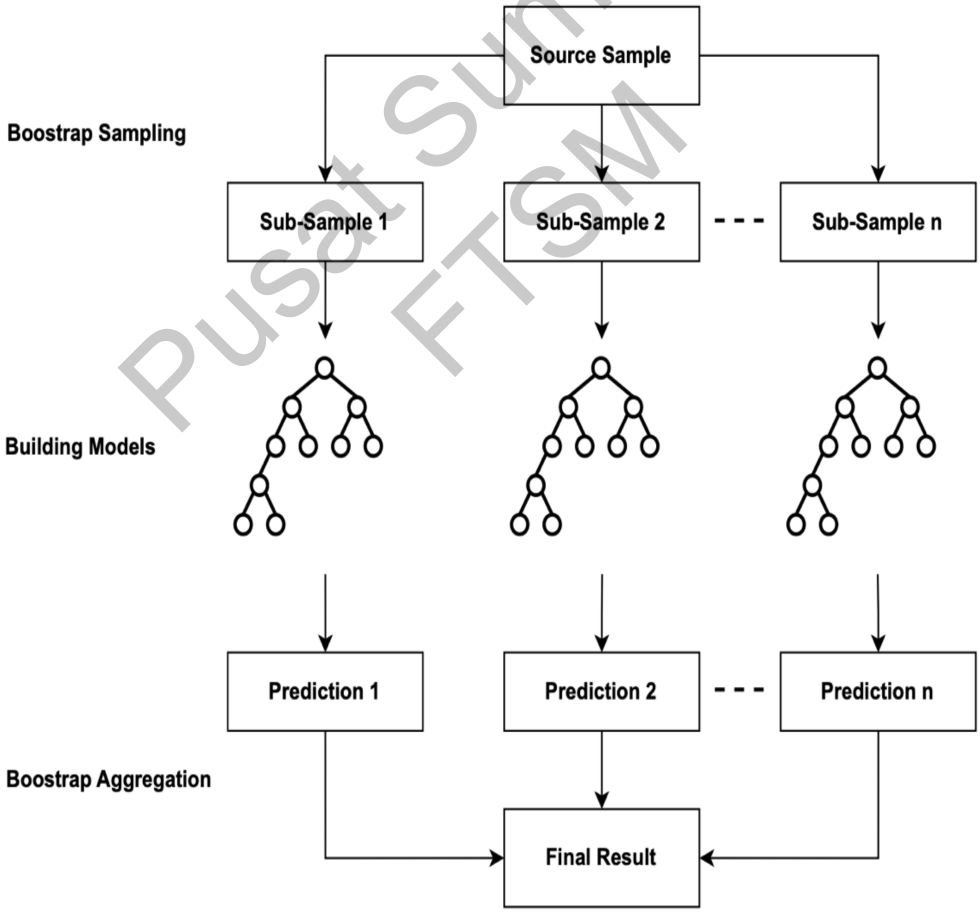

Figure 2.1 Random Forest Model Schema

Source: Sarkar & Natarajan 2019

2.4.3 Convolutional Neural Networks

Convolutional Neural Networks are highly efficient deep learning models specifically crafted for handling grid-structured data, like images. Key elements of a CNN comprise the convolutional layer, which executes multiple computational tasks, the max-pooling layer, tasked with compressing and smoothing data, and the fully connected layer, which merges all neurons to generate the final output in both the forward and backward passes (Mutabazi et al. 2023; Zulqarnain et al. 2021). Figure 2.2 depicts the structure of a CNN model.

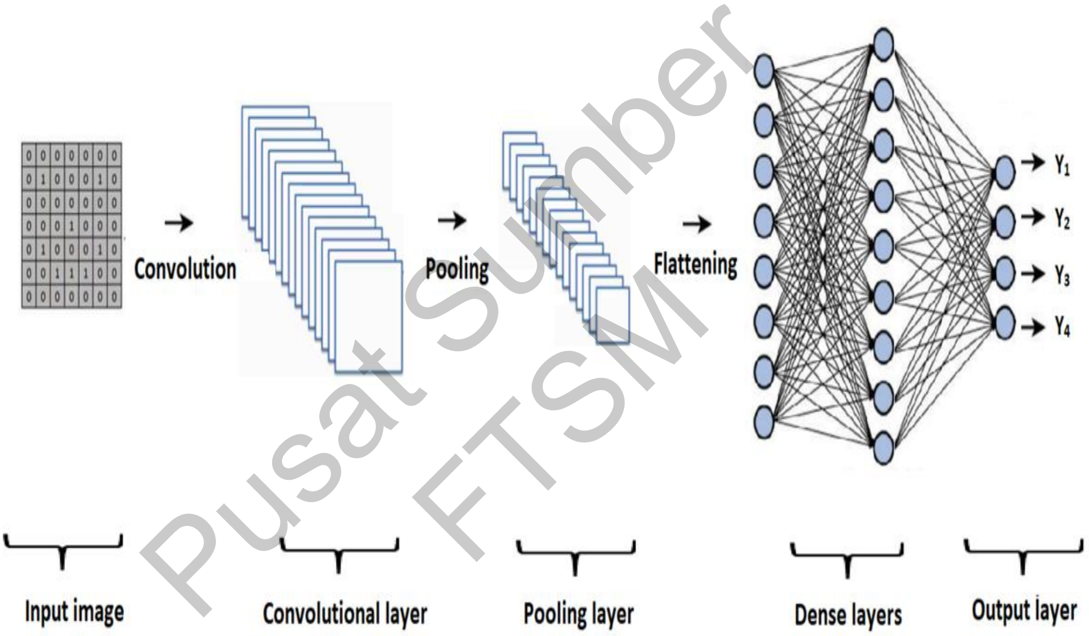

Figure 2.2 General Schema for CNN

Source: Maeda-Gutiérrez et al. 2020

For text data convolution operations, a wide convolution employs an n x d kernel, where n signifies the number of words captured per operation, and d indicates the word vector's dimensionality (Dachapally & Ramanam 2017). Varying kernel sizes relate to different n-grams: a kernel size of 2 targets lower-level word meanings (2-grams), whereas a kernel size of 5 captures higher-level information (5-grams) about the input (Seidakhmetov 2020).

2.5 FEATURE EXTRACTION TECHNIQUES

Feature extraction is a machine-learning process that transforms raw data into meaningful features. Identifying the most relevant and informative data improves the models' performance (Kowsari et al. 2019); this way, it is less likely to learn noise from the data and, therefore, reduce overfitting. This involves creating a list of words from the text data and converting it into features for classification purposes (Aninditya et al. 2019). Unigram is the simplest feature extraction method generated by forming a set of unique terms from the text data. Besides unigram, various other techniques are employed to derive feature sets, including bigram, trigram, and POS tagging (Gani et al. 2022). Previous studies such as (Aninditya et al., 2019 Mohammedid & Omar, 2020; Osman Gani et al., 2022) showcased the importance of feature extraction in improving machine learning models' performance. Based on the literature review, Bag of Words, TF-IDF, N-grams, and Word2vec have been extensively used to obtain features from the questions. Therefore, they will be used in the study and compared to select the most effective feature extraction technique for improving question classification accuracy.

2.5.1 N-grams

N-grams are statistical data constituting sequences of nth items derived from a text. It is beneficial in predicting what the following word in the sequence will be (Dake et al. 2023). N-grams are fundamental in natural language processing for text prediction, machine translation, and sentiment analysis tasks. There are five types of n-grams (unigram, bigram, trigram, 4gram, and 5gram), corresponding to the number of terms obtained from the text, with trigram being the most commonly used (Ko et al. 2012). For example, in the question: "What type of measurements can be performed using a vernier caliper" in terms of bigrams, features will be extracted such as ("what type", "type of", "of measurements", " measurements can", "can be", "be performed", "performed using", "using a", "a vernier". "vernier caliper").

2.5.2 Bag of Words

Bag-of-words (BoW) is an orderless feature extraction technique. BoW has been applied across various fields, such as NLP, computer vision, document classification, information retrieval, and spam filters (Kowsari et al. 2019). The BoW name comes from the representation of bags of words or bags of features of textual information (Qader et al. 2019). In a bag-of-words, the collection of text is thought of like a bag of words (Kowsari et al. 2019); for example, if question one is: “What type of measurements can be performed using a vernier caliper ” and question 2 is: “What can be accurately measured with a digital caliper”, then a group of words is produced for each question by tokenising the sentences to create a dictionary of the words, as follows:

1. Question one: “what”, “type”, “of”, “measurements”, “can”, “be”, “performed”, “using”, “a”, “vernier”, “caliper”
2. Question two: “what”, “can”, “be”, “accurately”, “measured”, “with”, “a”, “digital”, “caliper”

Each bag of words is represented as follows:

1. BoW 1: {“what:1”, “type:1”, “of:1”, “measurements:1”, “can:1”, “be:1”, “performed:1”, “using:1”, “a:1”, “vernier:1”, “caliper:1”}
2. BoW 2: {“what:1”, “can:1”, “be:1”, “accurately:1”, “measured:1”, “with:1”, “a:1”, “digital:1”, “caliper:1”}

In BoW, the order of the words is irrelevant, and the frequency of the word is its value. Now, assuming question 3 is “What type of measurements can be performed using a vernier caliper and a digital caliper”, then the Bag of words will be: {“what:1”, “type:1”, “of:1”, “measurements:1”, “can:1”, “be:1”, “performed:1”, “using:1”, “a:2”, “vernier:1”, “caliper:2”, “and:1”, “digital:1” }

2.5.3 Tf-idf

TF-IDF, which stands for Term Frequency-Inverse Document Frequency, is a widely used feature extraction method. It is a statistical technique employed to determine the importance of a word within a document relative to a collection of documents (Yang & Long 2023). It is highly used in many studies, such as (Anindityа et al. 2019; Mohammed & Omar 2020), because of its versatility and discriminative powers instead of a bag of words. TF-IDF integrates Term Frequency (TF) and Inverse Document Frequency (IDF). TF measures how frequently a word appears in a document. It is one of the most straightforward feature extraction methods used to evaluate the importance of a word in a single document. The formula of TF is given by:

$$\boldsymbol{TF(t,d) = \frac{C(t_d)}{T_d}} \qquad \ldots(2.4)$$

Where $C(t_d)$ is the number of times t appears in a document d, and T_d is the total number of terms in the document d. However, this method lacks global context and does not consider how common or rare the term is across the entire corpus (Gani et al. 2022). IDF measures how important a term is across a collection of documents by reducing the weight of terms frequently occurring in many documents and increasing the weight of rare terms. The formula of IDF is given by:

$$\boldsymbol{IDF(t) = 1 + \log\left(\frac{D}{d_t}\right)} \qquad ..(2.5)$$

Where D is the total number of documents in the corpus, and d_t is the number of documents containing the term t. TF-IDF is the multiplication of both TF and IDF (Gani et al. 2022).

$$\boldsymbol{TF - IDF(t,d) = TF(t,d)\ \times IDF} \qquad ..(2.6)$$

2.5.4 Word2vec

Word embedding techniques specialise in mapping words into N-dimension vectors of real numbers. Various word embedding methods have been proposed to translate unigrams into understandable input for machine learning algorithms such as Word2Vec, GloVe, and FastText (Kowsari et al. 2019). Word2Vec groups words with similar meanings into the same vector space. With enough data, word2vec can accurately predict a word's meanings based on its occurrence history (Styawati et al. 2022).

2.6 CONCLUSION

Machine learning fundamentals, text classification, and the application of question classification were discussed in this chapter. Furthermore, this chapter reviewed past studies on the application of machine learning for classifying questions into subject-specific chapters. Drawing from previous studies, it was evident that Convolutional Neural Networks consistently outperformed other models in various natural language processing tasks. Specifically, CNNs demonstrated remarkable performance in text classification. Logistic regression will be employed as a baseline model. Random Forest and Convolutional Neural Networks will also classify questions into pre-defined subject chapters.

CHAPTER III

METHODOLOGY

3.1 Introduction

In this chapter, we delve into the research methodology used to address the research questions and objectives outlined in Chapter 1. The research framework is based on CRISP-DM. This chapter will offer a thorough overview of the research planplan, the procedures used for data collection, data preparation, feature extraction, model development, and the primary metrics for assessing the model performance.

3.2 Overview of Research Methods

In this section, the research methods for this research are presented. The depiction of the research methods flow can be found in Figure 3.1. The model development for question classification based on subject-specific chapters commences with data collection. Questions will be extracted from two sources: IGCSE Physics prelabelled worksheets and past exam papers from (Pastpapers.co 2024) website. Questions obtained from past exams will be manually labelled, and both datasets will be concatenated. After this, data inspection will gauge the number of questions collected and identify duplicate entries. The following is data preparation, which consists of removing duplicate questions identified previously. A series of NLP techniques will be applied to prepare the dataset, which includes converting text into lowercase, removing unwanted symbols, tokenisation, removing stop words, removing non-words, and lemmatisation. The next phase involves extracting features such as Bag of Words, TF-IDF, N-grams, and Word2vec. The classification models are constructed using Logistic Regression, Random Forest, and Convolutional Neural Networks. The models will be assessed through performance metrics such as accuracy, recall, precision, and F1-score. Finally, to identify the best model for question classification based on subject-specific chapters, a statistical T-test will be performed. The detailed explanations of each step in the research methods will be discussed in sections 3.4 to 3.11.

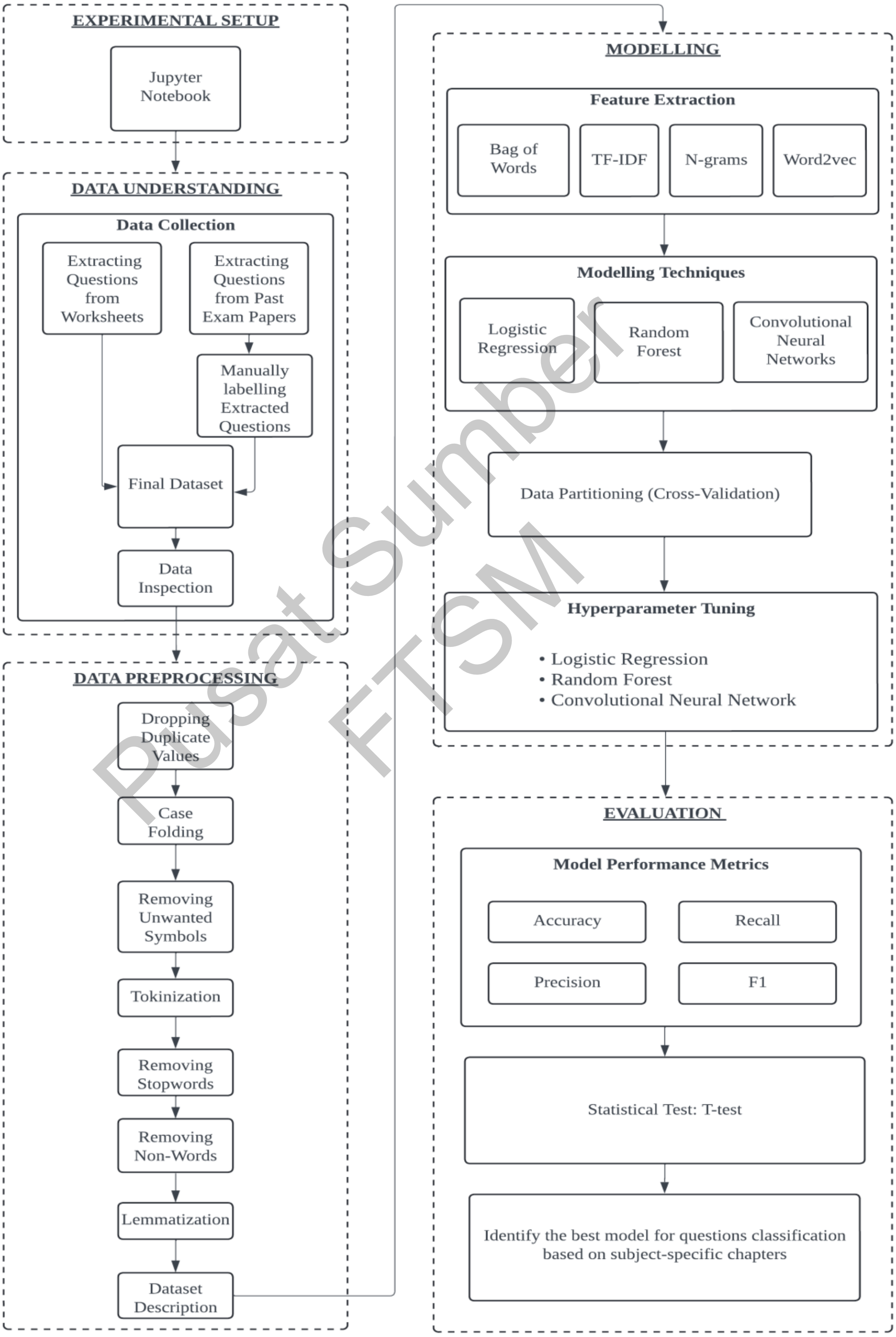

Figure 3.1 Overview of Research Methods for this Study

3.3 EXPERIMENT SETUP

This research was conducted using Jupyter notebooks. Jupyter Notebooks provides an interactive computing environment that supports various programming languages, including Python 3, which was used for this study. These notebooks are well-suited for machine learning and artificial intelligence tasks due to their ability to integrate code execution, text, and visualizations seamlessly (Jupyter 2024). For this study, the GPU was not used. This study used Python libraries such as Numpy, Pandas, PyPDF2, and CSV. These libraries were used to extract the questions and build the dataset. Matplotlib and Seaborn libraries were used for data visualisation. Furthermore, the NLTK library was used to preprocess the obtained dataset. Finally, Scikit-Learn and Keras were utilised for data transformation and model development.

3.4 DATA UNDERSTANDING

The study utilised data obtained from (Pastpapers.co 2024). This website provides learning materials for students, which include past exams and worksheets classified according to topics in PDF format. For this research, questions were extracted from both the pre-labelled worksheets and past exam papers. The past exam questions were manually labelled and concatenated to the labelled data. The methodology for extracting the questions will be further discussed in sections 3.4.1 and 3.4.2.

3.4.1 Questions from Worksheet

The dataset used for this research comprises 38 worksheets among 21 topics of five chapters from the IGCSE Physics subject. The average number of questions per worksheet is approximately 15 (14.89). The worksheets have a simple structure with one question per page, making the extraction process more manageable. To extract the questions, the PyPDF2 library, an open-source Python library capable of reading the pages of PDF files, was used. The page's contents are saved in a list data type variable to make manipulation easier. The header and footer are removed after changing the data type to string. Then, we split the text into parts using the question mark symbol (?). This makes the first part the question, while the other MCQ options. This is done because only the questions are used to train the model. Sometimes, the question does not have a question mark; thus, we split using “A.” since it indicates the first MCQ option. The following is an example of such a case:

Compared with β-particle and γ-rays, α-particles

A. are the sole radiation type carrying a charge

B. exhibit the most ionizing effect

C. possess the greatest penetrating power

D. possess the smallest mass

Next, multi-line questions are converted into a single line by replacing the new lines with a space (“ “), and extra spaces are removed. Finally, some questions that contain a picture or a diagram are not appropriately read by the Python library. PyPDF2 can only read string data; therefore, images cannot be parsed. Thus, an error text will be in the position of the image in the question. The unwanted error message will be removed, which concludes the question extraction process. This procedure is repeated for every page in all worksheets, saving the product as a new question in a final questions list.

3.4.2 Questions from Past Exam Papers

Exam papers from the 2019 IGCSE Physics summer exams, variants 21, 22, and 23, were used to obtain unlabelled past exam questions. Each variant contains 40 questions, and each page contains multiple questions, making the extraction process difficult compared to questions from worksheets. Similar to the previous methodology, the PyPDF2 library was used to obtain the contents of the pages. The contents were saved in a list data-type variable to simplify its manipulation. Headers and footers are then removed; however, it should be noted that the header and footer sizes are different for odd and even and for the last page of the exam paper. Therefore, the page number should be assessed before attempting to do so. This is important because the extraction procedure's core concept relies on having the first question number as the first item in the variable. Then, the MCQ options are discarded by removing lines that contain any of (“A ”, “B ”, “C ”, “D ”) after changing the data type into a string since these mark the MCQ options in the IGCSE exams and splitting the text with the question mark symbol (?).

With this, we should have a list variable containing multiple question items, and the first two alpha-numeric characters in the first item are the first question number on the page. Next, we strip the items from extra space and dispose of empty items in the list. Then, we should get the first question number on the page by obtaining the first two characters in the first item in the list and calculating the number of items. This is performed to know the possible question numbers on this page. Next, we split the items into a possible clean and unclean question. This is accomplished by checking if the first two characters in each item in the list are within the possible question numbers. Unclean questions are again split using the question mark symbol (?), and if an item starts with a potential question number, it is appended to the clean questions. Next, we check if the items in the list start with one number followed by two spaces or two numbers followed by two spaces. This is performed as these items are not questions but a part of another question. These items are removed, and we check for duplicate values in the list. Then, we check for uniqueness by checking if there is a question inside a question presented as another item in the list. Finally, we check if two items start with the same number and remove the shorter question. These questions were manually labelled and added to the questions from the worksheet.

3.5 DATA PREPROCESSING

Data preprocessing is an essential task in any text classification problem. The preprocessing process will clean the data from noise and improve the classification accuracy (HaCohen-Kerner et al. 2020; Kowsari et al. 2019). Various preprocessing methods include converting text from lowercase to uppercase, removing punctuation, or removing stopwords. More advanced techniques include lemmatisation, stemming, and part-of-speech tagging. Therefore, after obtaining the questions and the labelling, the next step is converting the unstructured data into a form appropriate for machine learning modelling. The following sections will discuss the preprocessing methods used for this study.

3.5.1 Dropping Duplicate Values

Dropping duplicate values is essential in preprocessing; we improve data quality and ensure data integrity. The classification model performance will increase and avoid overfiting. Since exam questions are sometimes reused between variants and worksheets, this step is an essential procedure in data preprocessing.

3.5.2 Case Folding

Sentences contain diverse usage of capitalisation, which makes it challenging to model. The machine learning model will interpret the same word in sentence case or uppercase as a different word. More importantly, word embedding techniques will struggle with the high feature space and the inability to link the same word with different casing. As such, the simplest way to resolve such problems is to convert all text into lowercase. This approach maps all words in text and documents to a common feature space (Kowsari et al. 2019).

3.5.3 Removal of Unwanted Symbols

Most text contains unnecessary symbols, such as punctuation. While crucial for human comprehension and interpretation, these symbols do not contribute meaningfully to machine processing. These symbols for this research include special characters such as Greek letters (α, β, etc.) and numbers. Therefore, these symbols have been removed for data preprocessing.

3.5.4 Tokenisation

Tokenisation is the procedure of segmenting a stream of text into words or characters. The text will be divided into words based on whitespaces for word-based tokenisation. The text is tokenised into individual characters or groups of characters for character-based (Kowsari et al. 2019). For this research, word-based tokenisation is applied using the NLTK library.

3.5.5 Stopwords Removal

Some words do not have significant value in text classification because they are present in most sentences. These words would only increase the feature space and reduce the model's performance. Such words include (“a”, “the”, “if”, and more); the best way to address such words is to remove them. For this research, stopwords were downloaded from the NLTK library and removed from each question.

3.5.6 Removal of Non-words

Due to the inability of the PyPDF2 library to read images properly, sometimes labels in images are extracted as text in place of the image position. The issue is that these labels are concatenated together; for example, the words “metal” and “rod” will be shown as “metalrod”. These words are usually already in the question, making their presence redundant. More importantly, splitting the words is very difficult; therefore, these words are removed. A word list that contains 466550 valid English words from (Mai 2024) is used to compare each word in the sentence. It will be removed if the word is not present in the words list.

3.5.7 Lemmatisation

Lemmatisation is a natural language processing technique that modifies the suffix of a word or eliminates it entirely to derive its base or root form. (Kowsari et al. 2019). For example, the word “walk” may appear differently throughout the text, such as walking or walked. Lemmatisation will strip the word to its original form (“walk”), reducing the feature space and improving the model's accuracy. For this research, lemmatisation is performed using WordNetLemmatizer from the NLTK library.

3.6 DATASET DESCRIPTION

Initially, the dataset comprised 566 questions from the worksheets and 120 from past exams. After dropping 32 duplicate values, the final dataset constitutes 654 unique questions, and the target variable is the chapters’ label. For Physics, there are five chapters; thus, the dataset has five target variables. The histogram in Figure 3.2 shows the class distribution of the acquired dataset.

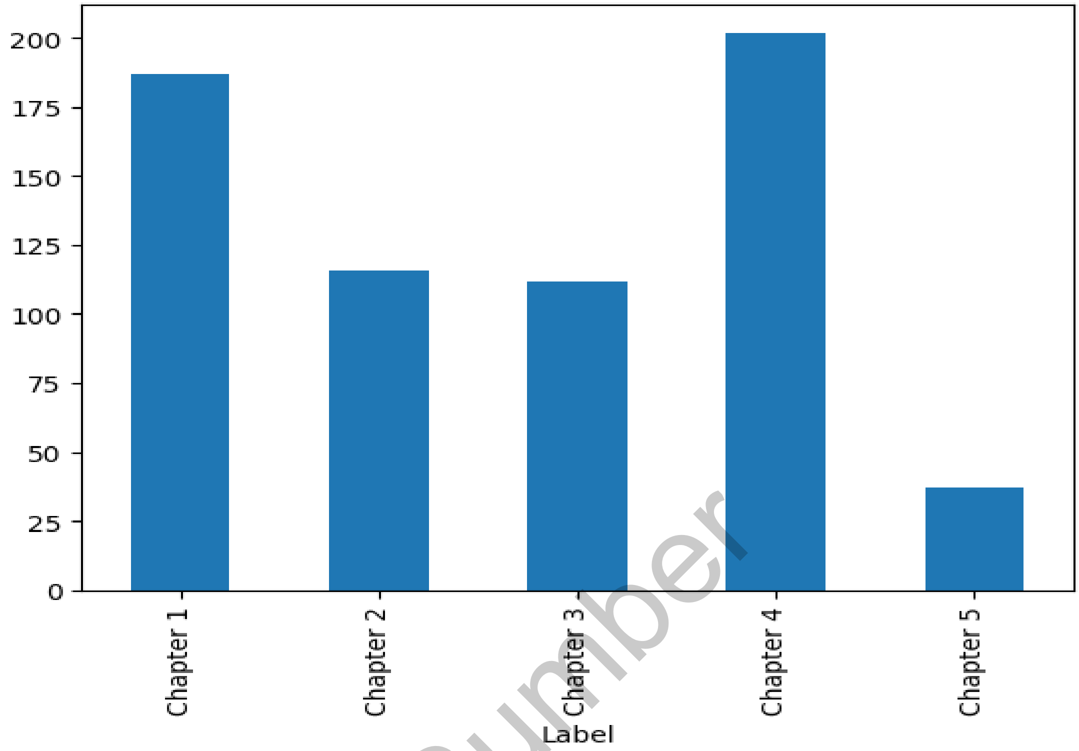

Figure 3.2 Class Distribution of the Dataset

3.7 FEATURE EXTRACTION

Feature extraction is an essential process in improving machine learning models. It involves converting raw data into meaningful features before classification. Methods such as TF-IDF and word2vec aid in prioritising terms and extracting semantic characteristics, enhancing the classification procedure (Liu 2022). In contrast, BoW focuses on representing text documents by counting word occurrences. On the other hand, n-grams provide statistical data by analysing the frequency of adjacent words. Previous studies have demonstrated that feature extraction techniques substantially influence classification outcomes, leading to improved and optimal results. This study uses methods such as BoW, TF-IDF, N-grams, and Word2vec to extract the features before modelling.

3.7.1 Bag of Words

Bag-of-words is among the simplest and most widely used feature extraction methods. This study used bag-of-words as a feature extraction method for Logistic Regression and Random Forest classification models. CountVectorizer() from the sklearn library was used to transform the corpus of questions into features. One thousand four hundred thirty-four features were obtained, and the top 10 features are listed in Table 3.1.

Table 3.1 Top Features using Bag-of-Words

Feature	Count
Show	238
Diagram	208
Water	116
Two	91
Wave	84
Wire	80
Force	79
Energy	77
Temperature	74
Circuit	73

3.7.2 Tf-Idf

TF-IDF gives greater weight to phrases that are significant in a particular document but are not frequent throughout the entire collection of documents (Aninditya et al. 2019). This methodology enables greater flexibility and accuracy in generating frequent item sets, enhancing classification accuracy rates (Gani et al. 2022). This study used TF- IDF in Logistic Regression and Random Forest through TfidfTransformer() from the sklearn library. The top 10 features are shown in table 3.2.

Table 3.2 Top Features using TF-IDF

Feature	TF-IDF
Show	24.0
Diagram	22.6
Wave	15.8
Water	14.7
Statement	14.4
Object	13.8
Circuit	13.7
Wire	13.0
Force	12.6
Energy	12.4

3.7.3 N-grams

N-grams aid in enhancing predictive modelling and classification accuracy by examining the frequency and co-occurrence of neighbouring words, enabling a greater understanding of the syntactic and semantic patterns found within the text. N-grams, such as bigrams and trigrams, offer crucial insights into the context and structure of the text, allowing for more precise feature extraction and classification (Aninditya et al. 2019). This study used unigram, unigram-bigram, bigram, bigram-trigram and trigram for feature extraction. The number of features derived from each n-gram is presented in Table 3.3.

Table 3.3 Number of Features for each N-gram

N-gram	Number of Features
Unigram	1434
Unigram-bigram	7557
Bigram	6123
Bigram-trigram	13383
Trigram	7260

3.7.4 Word2vec

Word2vec improves the performance of models by encoding words as vectors in space, leading to better outcomes in natural language processing tasks by detecting significant words based on their surrounding contexts (Zulqarnain et al. 2021). This approach utilises neural networks with a simple architecture, employing continuous bag-of-words (CBOW) and Skip-gram models. These models produce high-dimensional word vectors, effectively capturing semantic and syntactic relationships (Dachapally & Ramanam 2017). In addition, Word2vec considers subwords, which allows for generating new vectors for words that are not in the lexicon, enhancing subsequent tasks' performance (Luo et al. 2021). For this study, Word2vec was applied in the embedding layer in the CNN model using the genism library.

3.8 MACHINE LEARNING DEVELOPMENT

From the literature review conducted in Chapter 2, it was clear that Convolutional Neural Networks consistently achieved better results than other models in different NLP tasks. More precisely, CNN has shown exceptional performance in the text classification task. For this study, the baseline model will utilise logistic regression. Alongside LR, Random Forest and Convolutional Neural Networks will categorise questions into pre-defined subject chapters. The following sections outline the proposed architectures of the classification models.

3.8.1 Logistic Regression

The input for building the logistic regression model is the pre-processed set of questions, which comprises 654 questions categorised into five classes. Features were extracted from the questions using BoW, TF-IDF and N-grams and fed to the classifier independently. The parameter multi_class defaults to auto, which sets the classifier into multinomial classification since there are five classes. Therefore, the model uses cross-entropy loss to measure the effectiveness of a classification model. lbfgs optimisation Algorithm to be used by default along with l2 regularisation as a penalty to the loss function (Scikit Learn 2024a).

3.8.2 Random Forest

Like Logistic Regression, the Random Forest classifier takes the processed questions as inputs after feature extraction. During the initial phase, subdivisions of the dataset are generated using bootstrap sampling. The sample size of each subgroup is equal to the model input because the max_samples parameter is set to none by default. Each tree generates outputs representing the probability of belonging to one of five classifications. Initially, 100 trees were used for the random forest. The random forest classification model's final prediction is based on the target class's mean probability across all trees (Scikit Learn 2024b). The structure of the random forest model presented in this research is shown in Figure 3.2.

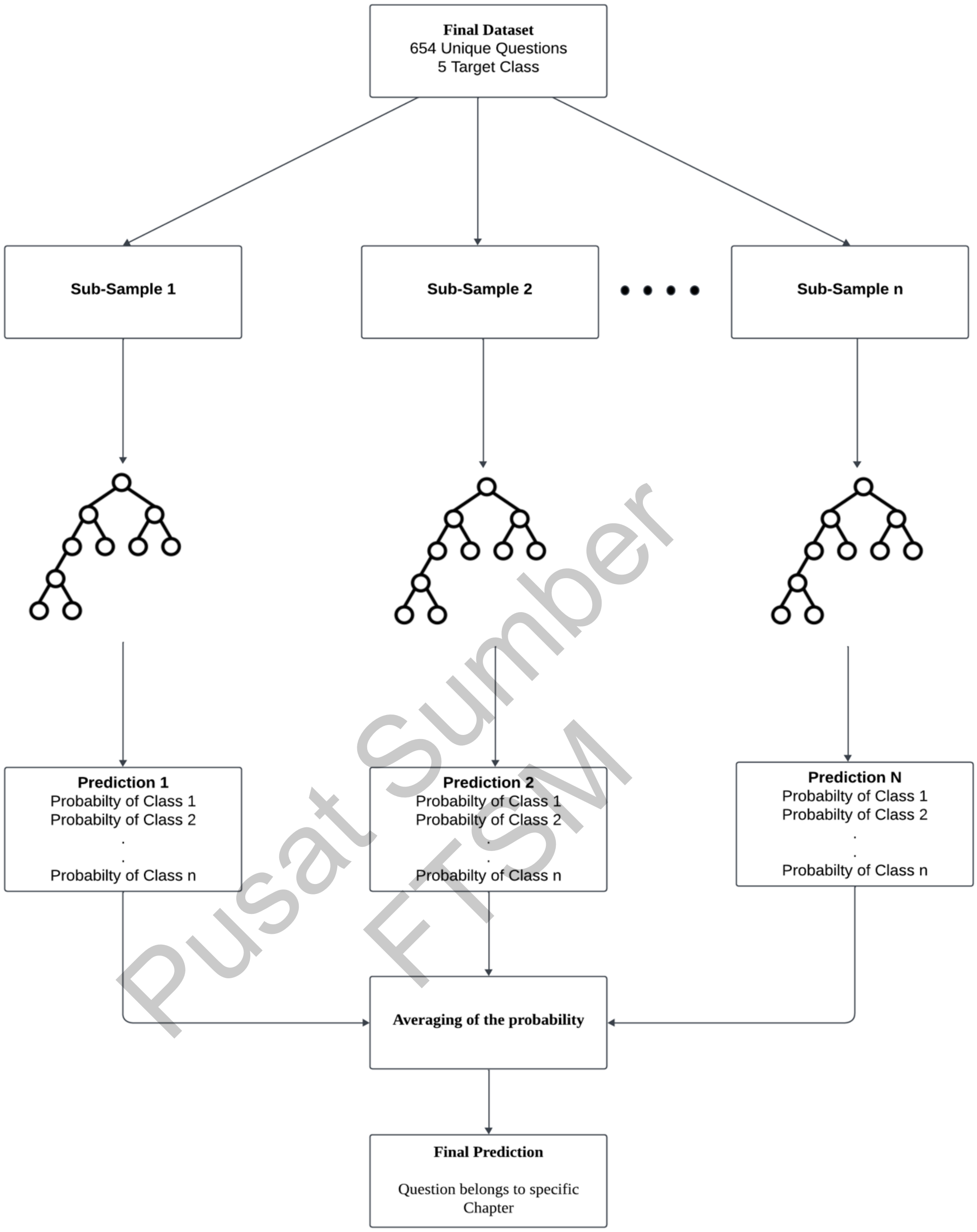

Figure 3.3 Structure of Random Forest for this Study

3.8.3 Convolutional Neural Network

Regarding Convolution Neural Networks, the preprocessed dataset undergoes a series of transformations to accommodate the CNN structure, starting with tokenising the corpus of text to define the vocabulary for the embedding layer and encoding the question words as integers. Next, the maximum length of input sequences (41) is obtained to pad all sequences to the fixed length. The vocabulary size for the embedding layer is set to 1434, which was taken from the previous BoW number of features. Finally, the text is encoded using the sklearn labelencoder() function. A standard question classification model uses an embedding layer as input, followed by a one-dimensional convolutional neural network, a pooling layer, and a prediction output layer. The embedding layer was set using genism, which builds the layer with weights from the Word2Vec model's learned word embeddings. The convolutional layer was set to 128 filters of size two and a relu activation function. Then, the Max Pooling layer consolidates the output from the convolutional layer—finally, a dense layer with a sigmoid activation function. The model was compiled with categorical cross-entropy and Adam optimiser. The batch size was set to 8, and the number of epochs was set to 20. The structure of the CNN model is shown in Figure 3.3.

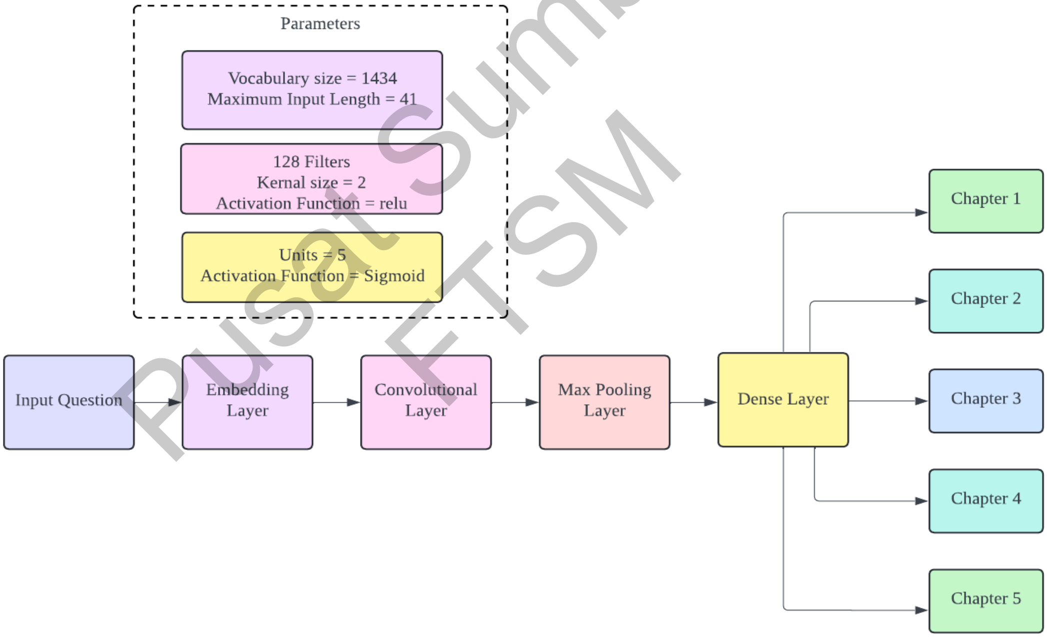

Figure 3.4 Structure of CNN Employed in this Study

3.9 DATA PARTITIONING (Cross-Validation)

Cross-validation is a technique used to assess the performance of machine learning models by partitioning the dataset into subsets for training and testing purposes (Dake et al. 2023). The process involves repeatedly training the model on a subset of the data and then evaluating it on the unseen data. This approach allows for multiple evaluations and helps prevent overfitting (Aninditya et al. 2019). The process is iterated numerous