Diagnosis for Patient and GP: Dialogue-based Self-Diagnosis with Disease-Symptoms Graph and Referral Letter Classification

by

Ruibin Wang

National Centre for Computer Animation

Faculty of Media & Communication

Bournemouth University

A thesis submitted in partial fulfilment of the requirements of Bournemouth University for the degree of *Doctor of Philosophy*

Oct. 2023

Acknowledgements

This thesis encapsulates the bulk of my research throughout my PhD journey—a remarkable odyssey, particularly since it coincided with the COVID-19 pandemic. With the world in chaos and high inflation adding to our burdens, life became even more challenging, yet we endured. The memories of this period will forever be etched in our minds because of our resilience in the face of adversity. I extend my heartfelt gratitude to everyone who cherishes peace and advocates for a regular life for ordinary individuals like you and me. Embarking on this PhD journey was both enlightening and demanding. I couldn't have reached this point without the invaluable help, encouragement, and support of countless individuals.

Firstly, I owe a deep debt of gratitude to my esteemed advisors, Professors Xiaosong Yang, Jian Chang, and Jianjun Zhang. I am particularly indebted to Professor Yang, whose expert guidance, unwavering positivity, and profound insights have been invaluable. Beyond his academic supervision, his life lessons and approach to facing challenges have deeply influenced my problem-solving perspectives. Our collaborative efforts in refining my papers, down to the minutest details, rank among my most treasured academic experiences.

I would like to extend my deepest gratitude to my peers and friends. A special mention goes to the senior PhD students, Dr. Jinglu Zhang, Dr. Yinyu Nie, Dr. Zhangcan Ding, Dr. Nan Xiang, Dr. Yao Lyu, Dr. Mengqing Huang, Dr. Shuangbu Wang, Dr. Yu Xia, and Dr. Yanran Li, for their invaluable suggestions and insights that added vibrancy to my life and studies in Bournemouth. The moments we spent cooking and celebrating together were instrumental in helping me navigate the challenges of the pandemic. I'm equally grateful to the current PhD students: Ms. Xiaoxiao Liu, Ms. Tingting Li, Mr. Junheng Fang, Mr. Kavisha Jayathunge, Mr. Abdul Rehman, and Mr. Jiajun Huang.

I wish them a prolific academic journey, filled with joy and a myriad of colourful experiences. A heartfelt thank you to Dr. Ehtaza, who graciously shared his office, providing me with a serene environment to complete my thesis. My sincere appreciation also goes to assistant professor Xiaokun Wang; our discussions on both academic subjects and life, our shared experiences watching football matches, engaging in sports, travelling, drinking, and cooking, truly enriched my life in the UK. Additionally, I'm forever thankful for the warmth and concern of my friends in China, and I hope they have good fortune and a happy life.

I extend my heartfelt gratitude to the faculty and staff of BU for cultivating a nurturing environment and for their invaluable guidance, advice, and resources. Additionally, I deeply appreciate the match-funded PhD scholarship from Bournemouth University and Shandong Hengdao Ruyi Digital Communication Ltd; without their generous financial support, this research would not have been conceivable.

A special note of thanks to my landlords, Yunfu Gong and Lihua You, for their kindness, care, and delightful dinners that sustained me during intensive project timelines. I hope for their continued health and happiness.

Finally, yet importantly, I want to express my profound gratitude to my family and relatives for their ceaseless love, encouragement, and steadfast belief in me. My parents, brother, and sister have instilled in me the values of diligence, perseverance, and the relentless pursuit of knowledge. To my girlfriend, for your unwavering understanding, love, companionship, sacrifices, and consistent support, I am forever grateful.

This thesis is submitted in fulfilment of the requirement for transfer from Master of Philosophy (MPhil) to the Doctor of Philosophy (PhD) at Bournemouth University.

Ruibin Wang

10 Oct. 2023

Abstract

The health systems worldwide face substantive operational challenges, more so after the exigencies precipitated by the COVID-19 pandemic, central to which is the pervasive shortage of medical resources at each health system level. This study explores two strategies aimed at alleviating these pressures by harnessing the potential of artificial intelligence (AI) and big data. The first strategy involves automated self-diagnosis tools for patients, which can help reduce the initial diagnostic burden on medical systems. The second strategy targets General Practitioners (GPs), aiming to bolster their diagnostic capabilities using AI-powered assistants.

Regarding the first strategy, current self-diagnosis tools face limitations in user interface and scope of diseases covered. In response, this study devised a dialogue-based self-diagnosis system. By harnessing data from the National Health Service (NHS) website, this thesis established a method to identify symptoms and generated a mapping of diseases to these symptoms, presenting as a disease-symptoms graph. Based on this graph, a two-choices diagnostic policy, statistically grounded, and devised respective solutions tailored to each module, are formulated utilising dialogue diagnosis strategies. Subsequently, a demonstration web application is created to showcase the dialogue-based diagnosis process. Concerning the second strategy, referral letters from GPs often contain rich diagnostic information, including medical histories, clinical observations, and initial diagnoses. Despite this, the use of these letters to aid diagnosis has been limited, primarily due to challenges in data labelling and anonymization. This study pioneers the use of referral letters as training data for disease diagnosis. However, data labelling and anonymization efforts are resource-intensive, resulting in a limited dataset size. As current methods proved inadequate for classifying diseases based on the gathered referral letters, this research proposes a hybrid architecture, which synergistically optimizes pre-trained encoder-based models and traditional deep learning models to fuse different representation spaces. It also innovatively introduces two data augmentation methods to underscore the importance of symptoms in the diagnostic process and to enhance feature representation performance. Our experiments showed that this approach significantly improved disease classification accuracy.

Additionally, the recent advancements in Large Language Models (LLMs) prompted us to explore their potential in analysing referral letters and decision-making. Specifically, the in-context learning performance of ChatGPT and GPT-4 in disease prediction is investigated. The results indicated that direct usage was suboptimal. Therefore, two disease classification fine-tuning solutions are proposed: supervised classification with encoder-based pre-trained language models (PLMs) and multiple-choice question-answering with LLMs. To address the challenge of limited training datasets, this thesis harnessed ChatGPT's text-generation capabilities to augment data effectively. The findings revealed that encoder-based models markedly surpassed decoder-based LLMs in disease classification from the augmented referral letters. Moreover, fine-tuning LLMs proved more effective than using GPT-4's few-shot learning. The experiment demonstrated that the optimal solution to assist GPs in clinical settings is combining the LLMs for data augmentation and the AI model based on encoder-based PLMs which achieve satisfactory performance for disease diagnosis.

List of Abbreviations

Artificial Intelligence (AI)
General Practitioners (GPs)
Intelligent Virtual Personal Assistant (Intel-PA)
Reinforcement Learning (RL)
Deep Q-network (DQN)
Natural Language Processing (NLP)
Convolutional Neural Network (CNN)
Long Short-Term Memory networks (LSTM)
Bidirectional Encoder Representations from Transformers (BERT)
Robustly Optimized BERT approach (RoBERTa)
Easy Data Augmentation (EDA)
Large Language Models (LLMs)
Name Entity Extraction (NER)
National Health Service (NHS)
Multiple-Choice Question and Answering (multiple-choice QA)
Naïve Bayes (NB)
K-Nearest Neighbors (KNN)
Support Vector Machine (SVM)
Bag-of-Words (BoW)
Term Frequency-Inverse Document Frequency (TF-IDF)
Continuous Bag of Words (CBOW)
Global Vectors (GloVe)
Recursive Neural Network (ReNN)
Multilayer Perceptron (MLP)
Recurrent Neural Network (RNN)

Convolutional Neural Network (CNN)
Generative Pre-trained Transformer (GPT)
Conditional Random Field (CRF)
Probabilistic Graphical Models (PGMs)
Logistic Regression (LR)
Naïve Bayes Transfer Classification (NBTC)
Expectation Maximisation (EM)
Hidden Markov Model (HMM)
Neighbour-Weighted K-nearest Neighbour (NWKNN)
Transductive Support Vector Machine (TSVM)
Decision Trees (DT)
Fast Decision-Tree (FDT)
Deep Neural Networks (DNNs)
Recursive Autoencoder (RAE)
Matrix-Vector Recursive Neural Network
Recursive Neural Tensor Network (RNTN)
Deep Recursive Neural Network (DeepReNN)
Paragraph Vector (Paragraph-Vec)
Bilateral Multi-perspective Matching (BiMPM)
Bidirectional LSTM (BiLSTM)
Deep Pyramid Convolutional Neural Network (DPCNN)
Transfer Capsule Network (TransCap)
Hierarchical Attention Network (HAN)
Question-Answering (QA)
Electronic Health Records (EHR)
Natural Language Understanding (NLU)
Dialogue Management (DM)
Natural Language Generation (NLG)
Dialogue State Tracking (DST)
Partially Observable Markov Decision Processes (POMDPs)
Semantically-Conditioned Generative Pre-Training (SC-GPT)
Knowledge-routed Deep Q-learning Network (KR-DQN)
Speech-to-Text (STT)

Text-to-Speech (TTS)
Google Text-to-Speech (gTTs)
Hospital Episode Statistics (HES)
University Hospital Dorset (UHD)
True Positives (TP)
True Negatives (TN)
Appointment Slot Issues (ASIs)
Named-Entity-Extraction (NER)
Reinforcement Learning from Human Feedback (RLHF)
Pre-trained Language Models (PLMs)
Parameter-Efficient Fine-Tuning (PEFT)
Low-Rank Adaptation (LoRA)
Knowledge Graph (KG)
Chain of Thought (COT)
Automatic Chain of Thought (Auto-CoT)

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The healthcare system has confronted significant operational challenges, intensified by the COVID-19 pandemic, notably the scarcity of medical resources across its various levels. While Artificial Intelligence (AI) and big data analytics have gained prominence, their use in healthcare remains relatively untapped. Properly harnessed, these technologies could alleviate strain across the health system. Two primary strategies have been identified to address this issue. The first focuses on patients, advocating for the development of automated self-diagnosis tools that provide preliminary diagnostic services, potentially reducing the demand on professional medical services. The second is oriented towards General Practitioners (GPs). Given that GPs often serve as the initial point of patient contact, enhancing their efficiency is crucial for timely healthcare delivery. Though GPs have broad medical knowledge, they might not have specialised expertise in every area. An AI-enabled assistant could bolster GP performance by providing specialised insights. This thesis will explore solutions and methodologies for both automated patient diagnosis and AI-assisted decision-making for GPs, aiming to alleviate the stresses on the healthcare system.

This research is embedded in the ongoing project titled "Intelligent Virtual Personal Assistant" (Intel-PA), which is a collaborative effort of a team. The development phase of Intel-PA involves notable collaborations with medical professionals from various institutions. Specifically, the self-diagnosis tools were developed in conjunction with physicians from Nuffield Health

Bournemouth Hospital, while the tools to assist GPs in decision-making were cultivated alongside doctors from University Hospital Dorset. My role within this project entailed translating the pragmatic requirements provided by the doctors into an academic framework and subsequently proposing academically oriented solutions. A detailed exposition of my contributions and methodologies employed within this project will be articulated in the subsequent sections of this thesis.

Owing to their affordability and pervasive accessibility, self-diagnosis tools are witnessing an escalating utilisation among adults for preliminary health assessments. These individuals are increasingly opting to employ automated diagnostic tools as an initial measure, prior to seeking professional medical consultation in a hospital setting. Consequently, the popularity of automatic diagnosis has experienced notable augmentation. Patient self-diagnosis can generally be conducted through three primary methodologies: online search-based, ontology-based, and dialogue-based diagnostic methods.

The prevalence of non-verified information online raises significant concerns, particularly in the context of search-based self-diagnosis services, which often yield irrelevant or, at times, nonsensical results, thus posing substantial risks to patients. The utilisation of ontology-based diagnostic methods (Lewenberg et al. 2017, Shim et al. 2018) demands that users engage in a meticulous process of form completion to ascertain symptom values, a task that can be laborious and test the patience of individuals, especially those who are unwell. This complexity may engender a lack of patience, potentially leading to the selection of inaccurate symptoms and subsequent deviations in diagnostic outcomes. In contrast, the method delineated by Lewenberg et al. (2017) portrays a typical clinical consultation wherein a patient initiates the dialogue by articulating feelings and symptoms through a self-report. Subsequent to this, the healthcare professional extracts explicit symptoms and proactively inquires about potential implicit symptoms, employing their medical expertise to establish a diagnosis through a multistep reasoning process. In each phase, the professional selects a pertinent symptom to explore further or concludes the diagnosis, considering both dialogue history and possible disease states. Furthermore, dialogue-based diagnostic tools, due to their simulation of real-world clinical consultation processes, are inherently user-friendly. Thus, dialogue-based self-diagnosis tools emerge as seemingly apt solutions.

In recent years, advancements in task-oriented dialogue systems have extended into the realm of intelligent medical diagnosis. Wei et al. (2018) pioneered the application of such systems to the medical field, formulating the dialogue processes as a Markov Decision Process and implementing a task-oriented medical dialogue system for disease determination via a Reinforcement Learning (RL) based approach. Such dialogue-based diagnostic systems predominantly depend on the intricate belief tracker—which estimates the patient's objective at every dialogue step—and data-driven learning. Direct application to automatic diagnosis is hindered by the paucity of training data. Xu et al. (2019) introduced a knowledge-routed relational dialogue system aimed at facilitating automatic diagnosis, employing a Deep Q-network (DQN) to navigate topic transitions (e.g., deciding which symptoms to inquire about) through data-driven learning while also considering the interrelations between diseases and symptoms. Nonetheless, the constrained scale of their training dataset—comprising merely 423 conversation data points—circumscribes their ability to glean pivotal insights into symptom-disease relations and optimally utilise medical knowledge. Consequently, their system restricts its diagnostic capability to a limited array of diseases. The development of novel methodologies to bolster the performance of dialogue-based diagnostic systems is imperative. Such methods should comprehensively encompass a wide array of prevalent diseases, ensuring both precise disease prediction and streamlined operational efficiency.

GPs exhibit proficiency in garnering critical patient information, encompassing medical history and present symptoms, thus crafting comprehensive referral letters for secondary care consultations. Despite referral letters containing rich medical information, due to the difficulties in handling unstructured format and stringent privacy regulations, their utility is constrained by their unstructured format and stringent privacy regulations. Besides, given the inherent challenges associated with referral letter data, such as the requisite for time-intensive and costly processing, medical professionals, typically constrained by time, often find data labelling to be an arduous task, limiting the availability of adequately prepared datasets. Despite these challenges and to our knowledge, there is a void in existing research concerning the leverage of referral letters to aid clinical decision-making processes. Consequently, this thesis pioneers an exploration into utilising referral letter data to enhance AI-assisted diagnostic decisions for GPs.

Given that these letters encapsulate rich, free-text information, Natural Language Processing (NLP) text classification techniques present a viable methodology for disease diagnosis. Nevertheless, the efficacy of NLP is substantively dependent upon the quality and volume of training data. The prevalent challenge encountered in this context is the scarcity of data, which hinders the development of a nuanced NLP model that can accurately generalise to unseen samples. Thus, addressing the insufficiency of training data for disease classification becomes a critical concern. Concurrently, formulating innovative, efficient, and accurate training architectures for the extant data is fundamental in propelling advancements in disease prediction through the analysis of referral letters.

In NLP research, disease diagnosis tasks can be principally addressed through two categories of models: 1) traditional deep learning-based text classification models and 2) pre-trained encoder-based models. The former encompasses Convolutional Neural Network (CNN) (Kim 2013, Conneau et al. 2016) and Long Short-Term Memory networks (LSTM) (Mousa and Schuller 2017), among others, whereas the latter involves, for instance, Transformer (Vaswani et al. 2017), Bidirectional Encoder Representations from Transformers (BERT) (Alaparthi and Mishra 2020), DistilBERT (Sanh et al. 2019), XLNet (Yang et al. 2019b), and Robustly Optimized BERT approach (RoBERTa) (Liu et al. 2019a). Both model categories present unique attributes and methodologies in processing linguistic data for diagnostic purposes, leveraging their respective architectures to discern and categorise relevant clinical information from text.

In contrast to conventional text classification tasks, the manifestation of symptoms within the content for classification proves integral to shaping the eventual outcome. The referral letters encompass not only verbose complaint texts but also succinct phrases that denote symptoms. Traditional deep learning-based models deploy static word vectors as features, derived from pre-trained word embedding models such as fastText (Kononenko 2001), word2vec (Xu et al. 2013, Kohavi et al. 1996, Nan et al. 2015), and GloVe (Hayashi 1990). These models tend to exhibit enhanced performance in the realm of short-text classification. The incorporation of pre-trained encoder-based models—serving as contextualised embedding models—offers the capability to extract long-distance dependencies in texts. This is achieved by utilising a multi-head attention approach to ascertain the similarities among words. Consequently, an amalgamation of these structures not only augments the representation of the neurology dataset but also reaps the benefits from both short-text symptoms and extensive complaint texts.

The challenges of labelling and anonymizing referral letters make the acquisition of real clinical data a formidable task. Thus, developing new data augmentation strategies becomes imperative to refine models and facilitate predictions for additional diseases. Existing data augmentation techniques, including Easy Data Augmentation (EDA) (Wei and Zou 2019), BERT (Alaparthi and Mishra 2020), BART (Lewis et al. 2019), and CBERT (Wu et al. 2019), enhance textual data by diversifying its representation. Nonetheless, the effectiveness of these methods may be limited due to the inherent restrictions in text representation and semantics.

The efficacy of Large Language Models (LLMs) like ChatGPT and GPT-4 (OpenAI 2023) has been validated in common sense examinations like the USMLE (Nori et al. 2023) and medical question-answering systems such as MultiMedQA (Singhal et al. 2022), where they exhibited remarkable performance. Additionally, their competence in Name Entity Extraction (NER) tasks on the NCBI (Doğan et al. 2014) and BC5CDR (Li et al. 2016a) datasets, as well as Relation Extraction (RE) tasks on GAD (Rouillard et al. 2016) and EU-ADR (Van Mulligen et al. 2012), has been evaluated, revealing unsatisfied results. Also the LLM platforms such as CancerGPT (Li et al. 2023), Med-PaLM (Singhal et al. 2022), and SynerGPT (Edwards et al. 2023) have demonstrated potential in specialised medical domains. However, their efficacy in authentic clinical contexts aiding GPs is yet to be established. Similarly, the potential of fine-tuning open-source LLMs such as LlaMA (Touvron et al. 2023), Aplaca (Taori et al. 2023), GLM (Zeng et al. 2022) for disease classification has not been investigated. Consequently, it is imperative to devise methodologies that scrutinise the efficacy and potential functions of LLMs within the realm of disease diagnosis

1.1 Motivation

The present healthcare system contends with significant operational challenges, particularly in the wake of the COVID-19 pandemic. To alleviate the strain on this system, collaborating with medical professionals from various institutions, this research proposes solutions oriented towards patient self-diagnosis and GP assistance.

Existing open-source

self-diagnosis tools manifest several limitations, including a dearth of credible data sources, an inefficient diagnostic process, and a restricted scope of disease coverage (Lewenberg et al. 2017, Shim et al. 2018, Wei et al. 2018, Xu et al. 2019). Consequently, there is an imperative to develop architectures that not only enhance the reliance and efficiency of patient self-diagnosis but also encompass a broad spectrum of prevalent diseases.

Referral letters are pivotal as they contain vital patient information, encapsulating both medical history and current symptoms. However, leveraging this data is hindered by challenges in handling unstructured formats and adhering to stringent privacy regulations, resulting in a dataset that is limited in size and utility. Consequently, its prospective role in assistant diagnosis is underexplored. This thesis pioneers the utilisation of referral letters for disease diagnosis employing NLP techniques. A primary concern emerges in addressing the deficiency of training data for disease classification, necessitating the proposition of data augmentation methods. Simultaneously, devising innovative, efficient, and accurate training architectures for the available data becomes instrumental to further advancements in disease prediction through referral letter analysis.

Furthermore, while large language models such as ChatGPT and GPT-4 have demonstrated substantial capabilities across various domains, their efficacy for disease diagnosis remains unverified. Thus, proposing a methodology to evaluate their performance in assisting GPs in disease diagnosis becomes a crucial research focus.

1.2 Research questions

In this section, the research questions of this thesis are introduced as follows:

- **Q1:** *The existing open-source self-diagnosis tools for patients cannot have credible data sources, efficient diagnostic processes, and large disease coverage at the same time. What methods can be used to solve this problem?*

- **Q2:** *In the case of a lack of referral letter datasets, how can features be extracted from a small amount of data for the auxiliary diagnosis?*

- **Q3:** *For low-scale data, the existing model performs un-satisfied, how to solve this problem?*

- **Q4:** *LLMs have a stunning performance on general tasks, but how do they perform in medical diagnosis? What methods can be used to verify their performance in clinical decision-making? How to use LLMs to enhance the performance of datasets or architecture for disease diagnosis?*

1.3 Aims and Objectives

The objective of this thesis is to introduce an efficient patient self-diagnosis architecture and to propose methodologies that support GPs in their decision-making processes. From an academic perspective, the primary challenge addressed herein involves innovating architectural models tailored for medical text classification, particularly when constrained by a limited training dataset.

1.4 Contributions

The contributions of this thesis are outlined as follows:

- A dialogue-based self-diagnosis system is presented. Data were extracted from the National Health Service (NHS) website, from which a methodology for symptom extraction was developed and utilized to establish a disease-symptoms graph. Subsequently, a two-choice diagnostic policy was proposed, grounded in statistical analysis. Solutions were correspondingly developed for each module, emphasising dialogue-based diagnosis. The culmination of this effort was the development of a demonstrative dialogue-based self-diagnosis web application.

- A hybrid architecture is introduced, amalgamating the features of words across diverse representation spaces, aiming to formulate primary diagnoses from referral letters. This architecture co-optimises the pre-trained encoder-based model and traditional deep learning model, thereby fully exploiting the two distinct representation spaces. Besides, Given the pivotal role symptoms occupy in disease diagnosis, two data augmentation methods: Complaint-Symptoms Integration Method and Symptom Dot Separating Method are proposed to amplify the performance of feature representation and underscore the significance of symptoms in the diagnostic process. In culmination, an AI diagnosis assistant web application, capitalising on the superior performance of this architecture, has been developed to assist GPs in rendering more accurate diagnoses.

- This research leverages the text-generating capacities of ChatGPT to augment referral letters through prompt engineering, utilising the enhanced dataset for refined disease classification. Consequently, this method enables the diagnosis of 17 distinct diseases, which is a notable improvement.

- This work proposed the direct diagnosis method and Multiple-Choice Question and Answering (multiple-choice QA) method to evaluate the zero/one/few-shot performance for disease prediction. Besides, using the ChatGPT augmented referral letter dataset, this work fine-tuned the performance of LlaMa/BERT/BioBERT on the disease classification, and the proposed pathway involves utilising ChatGPT for data augmentation, subsequently training encoder-based PLMs, and ultimately deploying the models to aid GPs in decision-making.

1.5 Outline of Thesis

This section will give a brief introduction of each chapter in this thesis as follows:

- **Chapter 1:** The research background, motivation, research problems, and main contributions of this thesis are introduced.
- **Chapter 2:** This research delves into AI diagnostic tools and decision-assistance mechanisms tailored for GPs, with a particular focus on medical text management, categorising our work within the scope of text classification tasks. This chapter initially elucidates the data processing procedure, followed by an introduction to fundamental models anchored in shallow learning, emphasising their pivotal role in enhancing text classification via sophisticated feature extraction and classifier design. Subsequently, this work navigates through the realms of traditional deep learning models as well as explores modern transformer-based architectures, both of which encompass multiple hidden neural network layers and exemplify a high degree of complexity, establishing them as widely utilised in contemporary research. This chapter serves as a repository of techniques, housing detailed evolutionary aspects of the methods deployed. The foundational theory underpinning the methodologies introduced from Chapter 3 through Chapter 5 is embedded within this chapter, facilitating a comprehensive understanding and resourceful reference point for readers.

- **Chapter 3:** The focus of this investigation centres on the systematic exploration of a dialogue-based automatic diagnosis system, meticulously addressing each module, including data collection, processing, graph construction, and the architectural design of diagnosis policies and the dialogue system. In response to the prevalent challenge of knowledge data scarcity, this chapter engineered a disease-symptoms graph, deriving data exclusively from the NHS website. Capitalising on the utility of the created disease-symptoms graph, the system adopted a knowledge-driven diagnostic policy, presenting a two-choice approach to manage the diagnostic procedure, accompanied by demonstrative examples within a subsequent web application platform.

- **Chapter 4:** This chapter elucidates the development and application of a hybrid architecture designed to enhance the precision of primary diagnoses conducted by GPs. Through meticulous comparative experiments utilising neurological datasets, it has been substantiated that the model proffers commendable classification accuracy. Additionally, two data augmentation methodologies are introduced: the Complaint-Symptoms Integration Method and Symptom Dot Separating Method, both of which have demonstrated efficacy in augmenting the predictive accuracy pertaining to neurological diseases. Furthermore, the hybrid architecture is extrapolated to forge an AI diagnostic assistant web application. This digital tool, capable of engaging in dialogues and text interactions with GPs, serves to adeptly assist them in executing preliminary diagnoses, thereby bridging technological advancements with practical medical utility.

- **Chapter 5:** In this chapter, the objective is to determine an optimal strategy that facilitates GPs in their decision-making processes by capitalising on referral letters and to evaluate the efficacy of LLMs in classifying diseases based on these letters. The direct application of LLMs for disease diagnosis was found to exhibit accuracy that is not commensurate with the requisite performance for tangible clinical scenarios. Nonetheless, findings illuminate a promising approach: employing ChatGPT for data augmentation, followed by the training of encoder-based PLMs, and ultimately implementing the refined models to aid GPs in their decision-making endeavours.

- **Chapter 6:** The conclusion of this thesis and the thought of developing directions for future work are provided.

1.6 List of Publications

- **Ruibin Wang**, Kavisha Jayathunge, Rupert Page, Hailing Li, Jianjun Zhang, Xiaosong Yang. *Hybrid Architecture Based Intelligent Diagnosis Assistant for GP*, BMC Med Inform Decis Mak 24, 15 (2024). https://doi.org/10.1186/s12911-023-02398-8.

- **Ruibin Wang**, Abdul Rehman, Tingting Li, Rupert Page, Hailing Li, Xiaokun Wang, Xiaosong Yang and Jianjun Zhang. *Capability of Large Language Models in Assisting GPs with Diagnoses.* Journal of the American Medical Informatics Association (under review).

- Xiaoxiao Liu, Eshani K. Fernando, **Ruibin Wang**, Mengqing Huang, Anthony Skene, Julia Ive, Jian Chang, Jian Jun Zhang. *Comparative Analysis of Synthetic Medical Datasets: A Study on Usability for Language Model Fine-Tuning.* Applied Sciences (Accepted).

1.6.1 Others

- Nan Xiang, **Ruibin Wang**, Tao Jiang, Li Wang, Yanran Li, Xiaosong Yang, and Jianjun Zhang. *Sketch-based modeling with a differentiable renderer*, Computer Animation and Virtual Worlds, 31.4-5 (2020): e1939.

- Shuangbu Wang, **Ruibin Wang**, Yu Xia, Zhenye Sun, Lihua You, and Jianjun Zhang. *Multi-objective aerodynamic optimization of high-speed train heads based on the PDE parametric modeling*, Structural and Multidisciplinary Optimization (2021): 1-20.

Chapter 2

Background

In the field of artificial intelligence for automatic diagnosis, the primary focus is on text-based tasks, which fall under the domain of NLP. This encompasses challenges such as text pre-processing, medical entity extraction, and diagnostic inference technology. Diagnostic inference could be categorised as a text classification problem. The techniques developed for text classification can be applied to medical diagnosis problems. Thus, this chapter will review the evolution of text classification methods and their applications in the medical domain.

Medical data typically comprises image and text components. Image data is processed using computer vision techniques, whereas text data is analysed through NLP methods. This thesis focuses on employing medical text data for automated diagnosis and AI-assisted decision-making, falling under the domain of text classification. Text classification involves assigning a given text to one or more predefined categories. This text can be as short as sentences, headlines, or product reviews, or as long as articles. The categories may include topics such as politics, sports, or military; emotional sentiments like positive and negative; ratings such as positive, neutral, or negative; or specific disease names. Such classifications can encompass both binary and multi-class challenges. Figure 2.1 outlines the stages of text classification, which encompass pre-processing, text classification methodology, evaluation, and labelling. This section delves into the components illustrated in Figure 2.1 and the intricacies of the methodological development.

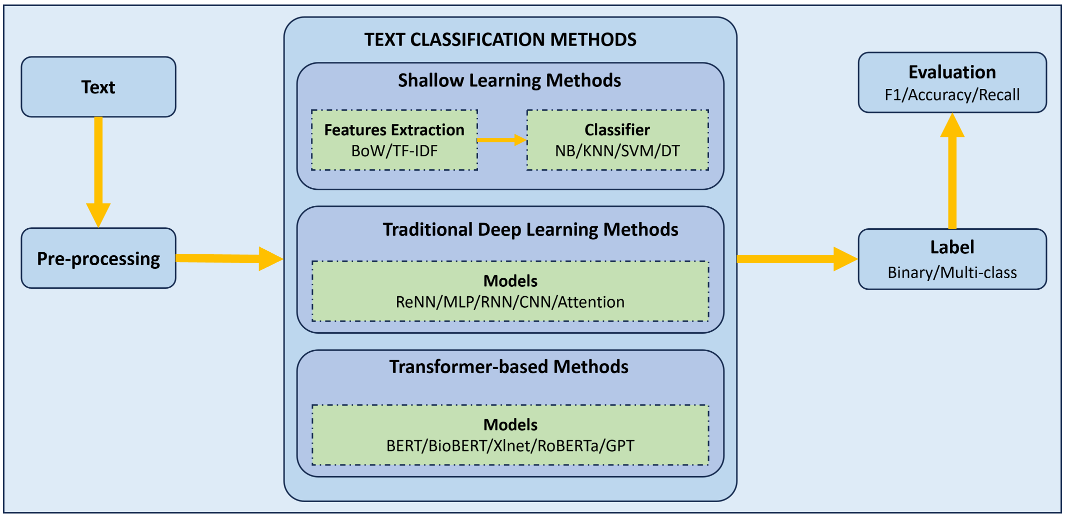


Figure 2.1: Various stages and their associated methods in text classification.

2.1 Method overview

During the early phases of machine learning, spanning from the 1960s to the 2010s, predominant text classification techniques were rooted in statistical methods such as Naïve Bayes (NB) (Maron 1961), K-Nearest Neighbors (KNN) (Cover and Hart 1967), and Support Vector Machine (SVM) (Joachims 1998). While these models provided noteworthy improvements in accuracy and stability compared to prior rule-based approaches, they relied heavily on manual feature extraction to optimise classification performance. This extraction process was both time-consuming and costly. Moreover, such models tended to overlook the sequence and context of textual information, potentially obscuring the nuanced meaning of words within sentences. By the 2010s, the realm of text classification started to shift towards deep learning models, with a marked transition to Transformer-based models around 2017. Unlike their predecessors, deep learning and Transformer-based models integrated the feature engineering step, learning the nonlinear mapping to directly connect features to outputs. This evolution not only eliminated the need for manual rule-setting and feature design but also facilitated more sophisticated semantic representations of text. It's worth noting that while

Transformer-based models are a subset of deep learning, their stellar performance and revolutionary impact merit distinct recognition. Current text classification research predominantly employs Transformer models, leveraging a plethora of pre-trained models and expansive language models to address a range of text-based challenges. Subsequent sections will delve into the modules pivotal to text classification and trace their evolution.

2.2 Text Pre-processing

In the medical domain, the text is often compiled by individual practitioners, leading to potential inconsistencies such as typographical errors and irregular characters. Consequently, text pre-processing is essential for effective feature extraction in text classification. This section presents methods for text data cleaning, aiming to eliminate inherent noise and enhance data characterization. Subsequently, two prevalent text feature extraction methods are discussed: weighted word representation and word embedding techniques.

2.2.1 Text Cleaning and Pre-processing

In text and document data processing, elements like stop words, misspellings, and slang frequently emerge. These elements may not influence classification outcomes but can interfere with certain classification algorithms, especially those based on statistical and probabilistic approaches. In the data processing endeavours elucidated in Chapters 3 through 5 of this thesis, these methodologies serve to augment the performance in the specific domain of medical text classification. Subsequent sections will review methods to refine and prepare text datasets.

- Tokenization (Gupta and Malhotra 2015, Verma et al. 2014)

  Tokenization plays an essential role in natural language processing as a pre-processing technique. It segments a continuous text stream into distinct units termed tokens, which can be words, phrases, symbols, or other meaningful elements. This segmentation facilitates the analysis and comprehension of the individual words within a sentence. Here gives an example of text tokenization:

Consider the sentence *"After a full day of study, he feels immensely satisfied."*

Upon tokenization, the resultant tokens would be:

["After", "a", "full", "day", "of", "study", "he", "feels", "immensely", "satisfied"].

- Stop Words (Saif et al. 2014)

  In the domain of text and document classification, some words serve primarily as connectors or modifiers without contributing substantial meaning. Such words, often referred to as "stop words," encompass terms like "a," "of," "again," "through," and "after." A common approach in classification algorithms involves the removal of these words from text datasets.

- Capitalization (Gupta et al. 2009)

  In text data analysis, varying capitalization can introduce ambiguity, leading to added complexity. A common solution is to convert all characters to lowercase. However, this method may unintentionally alter the meaning of certain words. For instance, transforming "US" (referring to the United States of America) to "us" (a pronoun) can result in misinterpretations. Implementing a slang and abbreviation converter can address these anomalies.

- Slang and Abbreviation

  In text pre-processing, addressing anomalies like slang and abbreviations is essential. Whitney and Evans (2010) defines an abbreviation as a shortened form of a word or phrase, often comprising the word's initial letters. For instance, "NLP" stands for "natural language processing". On the other hand, slang represents a unique linguistic category, appearing primarily in informal contexts, and frequently carries a meaning distinct from its standard counterpart. The phrase "lost the plot" for example, colloquially indicates that someone has become irrational. A prevalent method to handle such terms is their conversion to the standard language, as demonstrated by Dhuliawala et al. (2016). This standardisation is particularly vital in medicine, where patients' descriptions can deviate from established terminology. Hence, aligning these descriptions with recognised terms is imperative.

- Noise Removal

  In many text and document datasets, superfluous characters like punctuation and special symbols are prevalent. While these characters can be pivotal for human comprehension of documents, they may adversely affect classification algorithms.

- Spelling Correction

  Correction is vital in natural language processing, particularly for the dataset obtained from platforms like Twitter, where typos are frequent. A myriad of techniques has been proposed to tackle this challenge in the NLP domain. Mawardi et al. (2018) provides an overview of some of these methods. Notably, researchers can employ hashing-based and context-sensitive approaches (Dziadek et al. 2017). Additionally, Trie-based techniques and those using the Damerau–Levenshtein distance bigram have also been highlighted for their efficacy in spelling correction (Christanti et al. 2018).

- Stemming (Singh and Gupta 2016)

  In natural language processing, variations of a single word often share the same semantics, such as singular and plural forms of nouns. Stemming is a technique employed to standardise these diverse word forms into a consistent feature space. This method reduces a word to its base or root form by removing linguistic affixes. For instance, "caring" stems to "care."

- Lemmatization (Plisson et al. 2004, Korenius et al. 2004)

In natural language processing, lemmatization refers to the technique of obtaining the base form, or lemma, of a word by either replacing or removing its suffix. This process aids in standardising words to their canonical forms for better text analysis.

This section describes the techniques employed for text data pre-processing to mitigate potential noise from the source data that could adversely affect classification outcomes. The subsequent section will detail the approaches used to map the processed data into vector space.

2.2.2 Weighted Words

Weighted words, in the context of NLP and medical text analysis, typically refer to assigning values or weights to words in a text to signify their importance or relevance in a particular context. In the weighted word method, each document is transformed into a vector. The length of this vector corresponds to the document's length, and its elements represent the word frequencies within that document. Weighting words can be essential for various applications like text summarization, sentiment analysis, and information retrieval. The following section will review the methods used for weighted words.

2.2.2.1 Bag of Words

In the Bag-of-Words (BoW) model (Mikolov et al. 2013a), text is transformed into numerical vectors based on word occurrence. This representation prioritises word count over grammatical structure or sequential order. The BoW model finds applications in computer vision, NLP, Bayesian spam filtering, document classification, and machine learning-based information retrieval. Central to the BoW approach are three fundamental processes:

- Tokenization: This involves breaking down the text into individual words or tokens.
- Vocabulary Building: Here, a comprehensive vocabulary comprising all unique words from the dataset is constructed.

- Vectorization: Every document or text segment is converted into a vector. The vector's length is determined by the vocabulary size. Within this vector, the position of each word is designated according to its frequency in the respective document, or merely its presence, represented as a binary value (1 for presence, 0 for absence).

  The BoW model is recognised for its simplicity and efficacy across various applications. Its intuitive nature and ease of implementation make it particularly potent, especially when used in tandem with algorithms such as Naïve Bayes for tasks like text classification. However, the BoW model is not without its shortcomings. A prominent limitation of BoW is its insensitivity to word order. For example, the sentences "The dog stood on the desk" and "The desk stood on the dog" would have indistinguishable BoW representations, despite their differing meanings. Additionally, in the context of extensive vocabularies, the BoW model tends to produce sparse representations. These lead to elongated vectors predominantly filled with zero values, resulting in unnecessary computational overhead. Moreover, the model does not encapsulate semantic nuances. Words such as "happy" and "joyful," despite their contextual similarities, are treated as distinct entities.

  Here is a photophobia description example of BoW:

  Reports sensitivity to lights, particularly bright or flashing lights

  Bag-of-Words (BoW)

  ["Reports", "sensitivity", "to", "lights", "particularly", "bright", "or", "flashing"]

  The feature of this bag-of-words presented by frequency is [1, 1, 1, 2, 1, 1, 1, 1]

2.2.2.2 Term Frequency-Inverse Document Frequency

In the realm of text analysis, the Term Frequency-Inverse Document Frequency (TF-IDF) (Sparck Jones 1972, Jones 1973, Robertson 2004) emerges as a notable technique. This approach integrates the crucial elements of Term

Frequency (TF) and Inverse Document Frequency (IDF). The incorporation of the IDF component aims to reduce the influence of commonly occurring words in the corpus on the ultimate classification outcomes. Notably, IDF attributes a higher weight to terms that appear with significant frequency variation within a document. The expression for the weight of a term in a document, as determined by the TF-IDF method, is presented below:

$$W(d,t) = TF(d,t) * log(\frac{N}{df(t)}) \tag{2.1}$$

where N represents the total number of documents in the given context, while $df(t)$ represents the number of documents where term t is present. The first component of the equation $TF(d,t)$ is designed to enhance recall, while the latter $log(\frac{N}{df(t)})$ boosts precision in word embeddings. However, despite TF-IDF's efforts to mitigate the challenges posed by frequently occurring terms, it has inherent limitations. Notably, it fails to capture semantic similarities between words within a document, as each term is represented independently.

2.2.3 Syntactic Word Representation

Syntactic word representation refers to the representation of words in a manner that captures their syntactic roles and relationships within sentences. Syntactic information helps to understand how words combine to form phrases and sentences, ensuring grammatical correctness and conveying meaning through structure. Various methods have been developed to encapsulate syntactic information in word representations. Here N-Gram Bengio et al. (2000) and Syntactic N-Gram Sidorov et al. (2013) will be introduced.

2.2.3.1 N-Gram

In computational linguistics and natural language processing, the concept of an N-gram is very important Bengio et al. (2000). An N-gram refers to a sequence of n consecutive items from a given text or speech. For instance, when $n = 1$, it is termed a unigram (e.g., "apple"). For $n = 2$, it becomes a bigram (e.g., "apple pie"), and for $n = 3$, a trigram (e.g., "apple pie recipe"). As n increases, this section continues to identify sequences of corresponding lengths. One of the primary objectives of using N-grams is to encapsulate local word arrangements or structures in the text. While N-grams themselves aren't standalone representations of a text, they are instrumental as features to characterise it. An example of a simplistic representation is the BoW model, which utilises unigrams, thereby discarding word order. Despite its simplicity, the BOW model provides a concise vector representation of a text, typically proportional to the text's size. Employing N-grams, especially with $n > 1$, can capture richer contextual information than a mere unigram representation. To illustrate, consider a 3-Gram neurological examination example as follows:

Alert and oriented to person, place, and time. Cranial nerves II-XII are grossly intact. No noticeable motor or sensory deficits.

In this case, the tokens are as follows:

"Alert and oriented", "and oriented to", "oriented to person", "to person, place", "person, place, and", "place, and time.", "Cranial nerves II-XII", "nerves II-XII are", "II-XII are grossly", "are grossly intact.", "No noticeable motor", "motor or sensory", "or sensory deficits."

2.2.3.2 Syntactic N-Gram

In the study by Sidorov et al. (2013), syntactic n-grams were introduced. These differ from traditional n-grams, which are based on a text's linear structure; instead, syntactic n-grams derive from paths in a dependency or composition tree.

2.2.4 Word Embedding

In the field of language representation, relying solely on syntactic descriptions does not fully encapsulate the semantic nuances of words. For instance, in the bag-of-words model, there's a clear limitation in truly reflecting word semantics. Words like "tumour", "mass", "lump", and "lesion" might be contextually related, yet their vectors in this model remain orthogonal. Furthermore, this model fails to recognise the sequence of words within a phrase. Although n-grams aim to address this issue, they fall short and necessitate distinct similarity metrics for each word in a sentence. Khattak et al. (2019) reviewed the word embedding methods used in clinical text.

Word embedding transforms words or phrases into vectors within an N-dimensional real number space, serving as a robust feature extraction technique. This domain has seen significant research advancements, notably the contributions of Mikolov et al. (2013a) who introduced the Skip-gram and Continuous Bag of Words (CBOW) models. These models employ a single-layer architecture that leverages the inner product of word vectors. Techniques such as Word2Vec (Rong 2014), Global Vectors (GloVe) (Pennington et al. 2014), and FastText (Joulin et al. 2016) have gained recognition for their effectiveness in deep learning applications. A recent advancement in the field under discussion encompasses the emergence of "contextualised word representations." This concept signifies that vector representations of words exhibit variation in alignment with their contextual usage. Exemplar methodologies, including BERT (Devlin et al. 2018b), ELMo (Peters et al. 2017), and GPT (Radford et al. 2018), leverage this approach, and these will be elucidated upon in Section 2.3, wherein their application to text classification models will be discussed.

2.2.4.1 Word2Vec

The Word2Vec method employs a two-layer neural network to capture word semantics from vast text corpora. Once trained, the model produces a high-dimensional space, often spanning hundreds of dimensions, wherein each unique word from the corpus receives a specific vector. These vectors encapsulate the semantic characteristics of words, making them suitable for a variety of NLP applications. Word2Vec encompasses two main architectures: CBOW and Skip-gram. Notably, the Skip-gram architecture evaluates a corpus using target words w and their respective context c (Mikolov et al. 2013a b). The primary objective of this methodology is to enhance the correlation probabilities as illustrated:

$$arg \max_{\theta} \prod_{w \in T} [\prod_{c \in c(w)} p(c \mid w; \theta)] \tag{2.2}$$

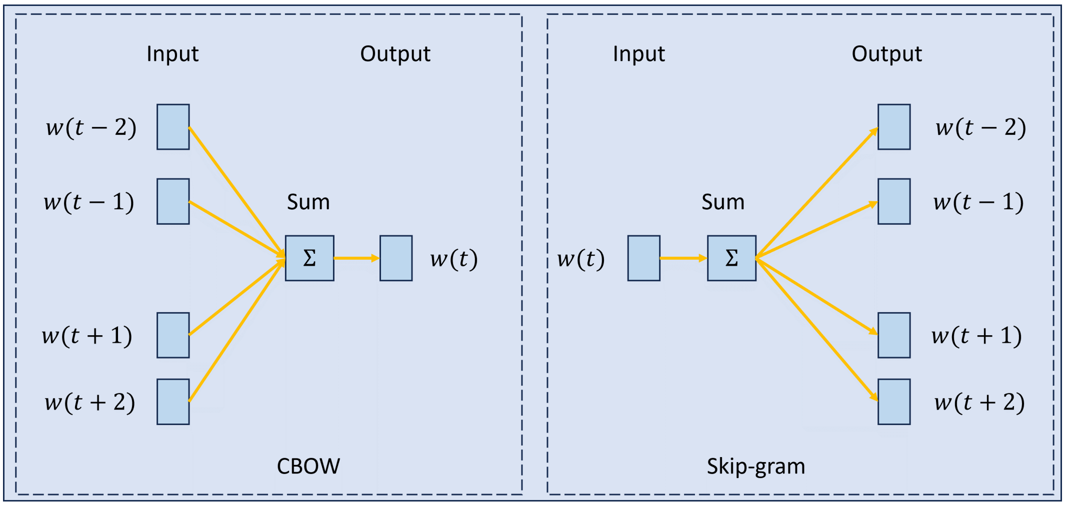


Figure 2.2: The architecture of CBoW (left) and Skip-gram (right)

where T refers to Text, and θ is parameter of $p(c \mid w; \theta)$.

In Figure 2.2, the left panel illustrates the CBOW model, which predicts the target word using its surrounding context words. Conversely, the Skip-gram model, displayed in the right panel, determines the context words when provided with a specific target word.

Continuous Bag-of-Words Model

In natural language processing, the CBOW model is notable for its unique strategy of predicting a target word using its surrounding context words. Consider the word "war" as an example: "tank" and "military" might serve as context words in a CBOW context. This approach necessitates duplicating the connection input to the hidden layer β times, with β denoting the number of context words (Mikolov et al. 2013a). As a result, the CBOW model vectors represent unordered word sets. To initiate this process, a vocabulary comprising all unique words in the corpus is constructed. The subsequent output from the associated shallow neural network predicts words from their context. Notably, the number of context words is typically ranging from 4 to 5 words.

Continuous Skip-Gram Model

Recent developments highlight the potential of CBOW models to capture both syntactic and semantic information in sentences for machine learning algorithms (Mikolov et al. 2013a). A related architecture, the continuous skip-gram model, adopts a distinct strategy. Unlike CBOW, which predicts a word from its surrounding context, skip-gram optimises the classification of a word relative to another word in the same sentence. Both models are pivotal in maintaining intricate word relationships within sentences.

2.2.4.2 Global Vectors for Word Representation

In the realm of text classification, the Global Vectors (GloVe) (Pennington et al. 2014) embedding technique stands as a noteworthy method. Drawing parallels with the Word2Vec technique, GloVe represents each word with a high-dimensional vector. This representation is refined by examining the context provided by its neighbouring words within a vast corpus. A prominent pre-trained embedding that finds frequent application stems from a vocabulary of 400,000 words. This model was trained on a corpus consisting of Wikipedia 2014 data and Gigaword 5, resulting in vectors with a 50-dimensional representation for each word.

GloVe further offers a variety of pre-trained embeddings with higher dimensions, namely 100, 200, and 300. These have been trained on even more expansive corpora, encompassing content from platforms like Twitter. To better understand the spatial relationships between these embeddings, Figure 2.3 provides visualised examples of Global Vectors for Word Representation using the t-SNE technique (Van der Maaten and Hinton 2008). The mathematical foundation of GloVe to calculate the relationship between two words (for example i and j) is examined by finding co-occurrence probability with some probe words k, which is described by the following objective function:

$$F(w_i, w_j, \widetilde{w_k}) = \frac{P_{ik}}{P_{jk}} \tag{2.3}$$

$$P_{ik} = P(k|i) = \frac{x_{ik}}{Xi} \tag{2.4}$$

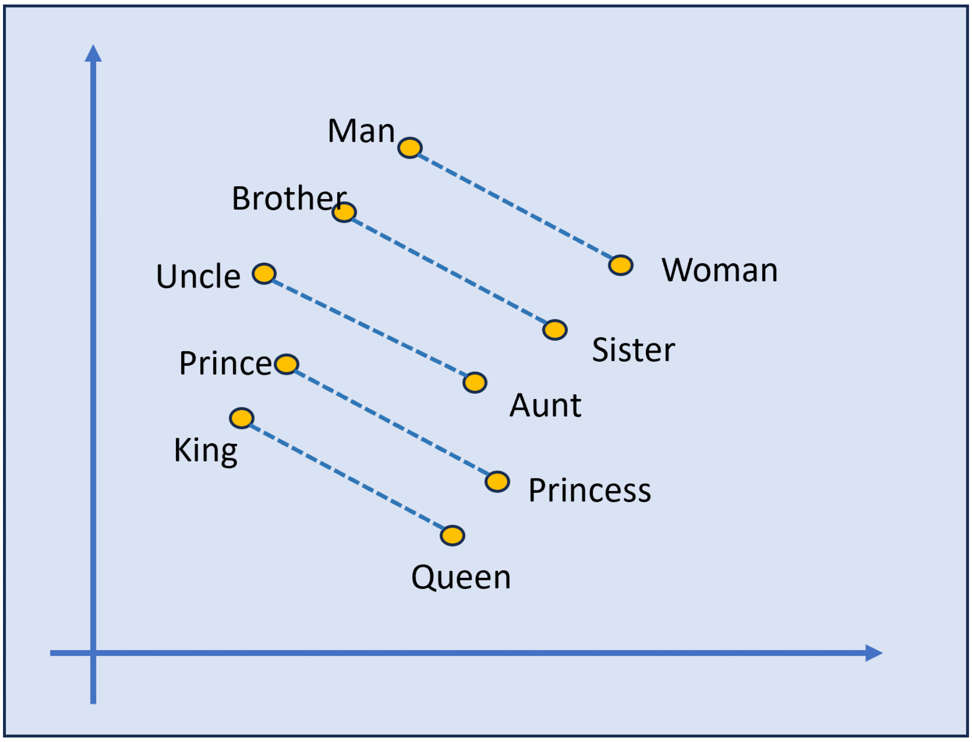


Figure 2.3: Examples of Global Vectors for Word Representation.

$$X_i = \sum_m X_{im} \tag{2.5}$$

where w_i, w_j and $\widetilde{w_k}$ are vectors for the word i, j and k, respectively. X_i is a summation of all the words which occur in the context of word i, and m is the length of X_i.

Consider two-word vectors: w_i represented by 'Ice' and w_j as 'Steam'. Taking probe words vector $\widetilde{w_k}$ such as 'Solid', 'Gas', 'Water', and 'Fashion', it can be inferred the following: 'Solid' relates more to 'Ice', 'Gas' is closely associated with 'Steam', 'Fashion' has no connection to either, and 'Water' relates to both 'Ice' and 'Steam'.

2.2.4.3 FastText

Morphology, in the context of linguistics, is the study of the structure and formation of words. It explores how words are constructed from morphemes, which are the smallest units of meaning in a language. It is often overlooked in many methods (Bojanowski et al. 2017), resulting in unique vectors being allocated to each word in most conventional word embedding models. To address this limitation, Facebook AI Research introduced FastText (Joulin et al. 2016), a pioneering word embedding method. In this approach, a word is characterised by a collection of character n-grams. To illustrate, for the word "embedding" with $n = 3$, FastText generates a representation using character tri-grams as follows:

$$< em, mbe, bed, edd, ddi, din, ing, ng >$$

Consider the sequence $< bed >$, which denotes a distinct word. It is crucial to differentiate this from the tri-gram *"bed"* derived from the word *"embedding"*.

In a recent publication, Facebook introduced pre-trained word vectors for an extensive array of 294 languages. These vectors were diligently trained on Wikipedia datasets utilising the FastText methodology. Notably, the training adopted a 300-dimensional representation and hinged on the Skip-gram model, with all parameters set to their defaults.

Summary This section entails the cleaning and pre-processing of raw input text to prime it for word representation models. Subsequently, the objective of text representation is to translate this preprocessed text into formats comprehensible for computers while mitigating information loss. Several techniques serve this purpose: BOW (Harris 1954) encapsulates each text as a vector, sized to the dictionary, with each value signifying the word's frequency in relation to its position. N-gram (Cavnar et al. 1994) expands upon this by factoring in adjacent words to construct its dictionary. TF-IDF (Hinterberger et al. 2009) models text by leveraging word frequency against the inverse of its document frequency. Techniques like word2vec (Mikolov et al. 2013a) harness local context for generating word vectors, while GloVe (Pennington et al. 2014) incorporates both local context and overarching statistical attributes, training on non-zero components of a word co-occurrence matrix.

2.3 Text Classification Methods

Text classification entails the extraction of features from text data to predict their respective categories. Over the past decades, a plethora of models have been developed for this purpose. These models find applications across various domains such as sentiment analysis, news or topic classification, question answering, natural language inference, disease prediction and event prediction. The previous section introduced methods for representing words as vectors. Now, we'll review techniques to extract features between words and complete the classification task using these vectors. We'll cover three types of methods: shallow learning, deep learning, and transformer-based. These will be detailed in the sections that follow.

As for deep learning, Recursive Neural Network (ReNN), Multilayer Perceptron (MLP), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN) and Attention-based Method are introduced. In these methods, TextCNN (Kim 2014) stands out with a substantial citation count. It brought forth the use of the CNN model in addressing text classification challenges for the first time. While Transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018a), Generative Pre-trained Transformer (GPT) (Floridi and Chiriatti 2020), XLnet (Yang et al. 2019b), and Robustly Optimised BERT Approach (Roberta) (Liu et al. 2019a), weren't initially tailored for text classification, their effectiveness across multiple datasets has led to their widespread adoption in text classification model design. Mujtaba et al. (2019) provided a comprehensive review of clinical text classification methods available up to that year, establishing a foundational context for the exploration of methodologies in this domain. Moreover, the application of pre-trained models and the practice of prompt tuning for the purpose of clinical text classification are meticulously explored in the review by Wang et al. (2023b), offering insights into more recent advancements and applications in the field.

2.3.1 Shallow learning

In the realm of text classification, compared with rule-based methods (Qin et al. 2009, Lawrence and Wright 2001), shallow learning models have broadened their application scope owing to their ability to deliver enhanced accuracy efficiently. As for shallow learning, the Naïve Bayes model (Maron 1961), marked the onset of text classification. Subsequently, general-purpose classifiers like K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Conditional Random Field (CRF) (Lafferty et al. 2001) emerged and became widely adopted in text classification tasks. In both comparative experiments and direct applications, these methods have been employed in the field of medical diagnosis. This includes predicting heart disease (Keerthana 2017, Bahani et al. 2021, Sarkar and Koehler 2012), cardiovascular risk (Bandyopadhyay et al. 2015), and in pregent diagnosis and treatment (Song et al. 2021), and among others (Andreassen et al. 1999, Verduijn et al. 2007, Lipsky and Lewis 2005, Vila-Francés et al. 2013, Lappenschaar et al. 2013). These methods are also used to guide the selection of the appropriate treatment (Lucas et al. 2000, Kazmierska and Malicki 2008, Yet et al. 2013, Velikova et al. 2014) and implemented as part of clinical decision support systems (Lucas et al. 1998, Sesen et al. 2013). This section will introduce these methods.

2.3.1.1 PGM-based methods

In Probabilistic Graphical Models (PGMs), conditional dependencies between features are encapsulated using graph structures. Examples include the Bayesian network (Zhang and Zhang 2010), hidden Markov networks (Eddy 1996), Logistic Regression (LR), and Conditional Random Field (CRF) (Lafferty et al. 2001). These models blend principles from both probability and graph theory. Figure 2.4 illustrates their structural composition.

Naïve Bayes (NB): In the realm of probabilistic models, the Naïve Bayes (NB) algorithm stands out for its simplicity and widespread application. Originally introduced by Maron (1961), this method is rooted in the application of Bayes' theorem. A fundamental assumption underpinning the NB approach is the conditional independence of features $\mathbf{x} = [x_1, x_2, \ldots, x_N]$

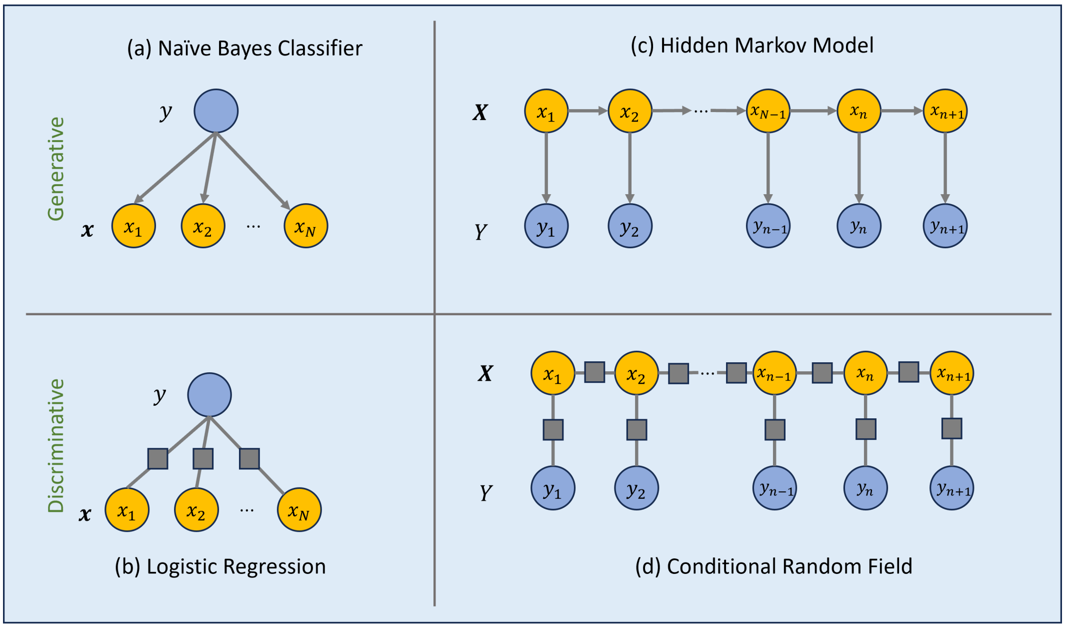


Figure 2.4: Probabilistic graphical models.

given a target value y, as depicted in the Figure 2.4(a). Central to its functioning, the algorithm harnesses prior probabilities to compute the posterior probabilities.

$$p(y, \mathbf{x}) = p(y) \prod_{n=1}^{N} p(x_n|y) \tag{2.6}$$

In text classification tasks, the NB algorithm is frequently chosen due to its straightforward structure. Notably, its underlying assumption of feature independence, though not always accurate, simplifies computations and often enhances performance. Schneider (2004) introduces a feature selection score methodology based on KL-divergence to boost performance for smaller categories in multinomial NB text classification. Similarly, Dai et al. (2007) put forth the Naïve Bayes Transfer Classification (NBTC), a transfer learning method, to bridge discrepancies in distributions between training and target sets. This technique employs the Expectation Maximisation (EM) algorithm (Dempster et al. 1977) to derive a locally optimal posterior hypothesis for the target set.

Logistic Regression (LR): In a technical context, Logistic Regression (LR) which is shown in Figure 2.4(b), serves as a statistical technique designed to forecast a binary outcome predicated upon multiple predictor variables. This method estimates the likelihood of a particular event or class occurrence, exemplified by outcomes like pass versus fail or alive versus dead. Notably, LR outputs a value ranging from 0 to 1, derived through the employment of the logistic (often referred to as the *sigmoid*) function.

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+...+\beta_N x_N)}} \tag{2.7}$$

Given the predictors X, the probability of class 1 is denoted by $P(Y = 1|X)$. The function constrains its outputs to lie within the interval $[0, 1]$. The model employs parameters $\beta_0, \beta_1, ..., \beta_N$.

Hidden Markov Model (HMM) (Eddy 1996) is a statistical model that represents a Markov process with hidden states. Particularly apt for sequential text data, HMMs enhance algorithmic efficiency through an optimised model structure. The model functions on the premise that there is an underlying process, denoted as X, and an observable process Y that is dependent on X. The objective of the learning process is to deduce the nature of X by monitoring Y, taking into account the interdependencies between states (refer to Figure 2.4(c)). Recognising the importance of contextual information in texts, Frasconi et al. (2002) transformed texts into sequences of pages to harness the inherent sequential relationships for multi-page text analysis. However, such techniques demonstrated limited efficacy in domain-specific texts. To address this, Yi and Beheshti (2009) incorporated specialised prior knowledge, predominantly sourced from the Medical Subject Headings (MeSH) vocabulary set (Salaba and Chan 2023), to enhance medical text classification.

Conditional Random Field (CRF), illustrated in Figure 2.4(d), is an undirected graphical model designed to harness the strengths of both classification and graphical modelling. CRFs uniquely merge the capacity to represent multivariate data succinctly with the capability to utilise a high-dimensional feature space, particularly beneficial for text data. Instead of modelling the joint probability $P(X,Y)$, CRFs focus on the conditional probability of a label sequence Y given an observation sequence X, denoted as $P(Y|X)$. This approach allows the inclusion of intricate features in the observation sequence without contravening the independence assumption (Sutton et al. 2012, Vail et al. 2007, Chen et al. 2017). However, the most pronounced challenge associated with CRFs is the considerable computational demand during the training phase, exacerbated when handling text datasets due to their expansive feature space. Moreover, the model struggles with words not encountered in the training set (Tseng et al. 2005).

2.3.1.2 KNN-based method

The K-Nearest Neighbors (KNN) algorithm, as described by Cover and Hart (1967), assigns a category to an unlabeled sample by evaluating the most frequent category among its k nearest labeled samples. This method, depicted in Figure 2.5(a), classifies a specific text based on its proximity to k training texts, exemplified for both $k = 3$ and $k = 9$. A primary advantage of KNN lies in its direct classification without the requisite model construction, thus offering streamlined complexity. However, its efficiency diminishes with substantial datasets due to the direct correlation between the algorithm's time and space complexity and data volume. Addressing the potential challenge of feature overload, Soucy and Mineau (2001) introduced a variant of the KNN algorithm that omits feature weighting. This version pinpoints pertinent features by establishing word interdependencies via feature selection. A noted limitation of the conventional KNN is its proclivity to favour classifications with more abundant data, especially in imbalanced datasets. To rectify this, the Neighbour-Weighted K-nearest Neighbour (NWKNN) algorithm (Tan 2005) was introduced. NWKNN refines classification accuracy in uneven datasets by assigning higher weights to neighbours from less-represented categories and lower weights to those from prevalent ones.

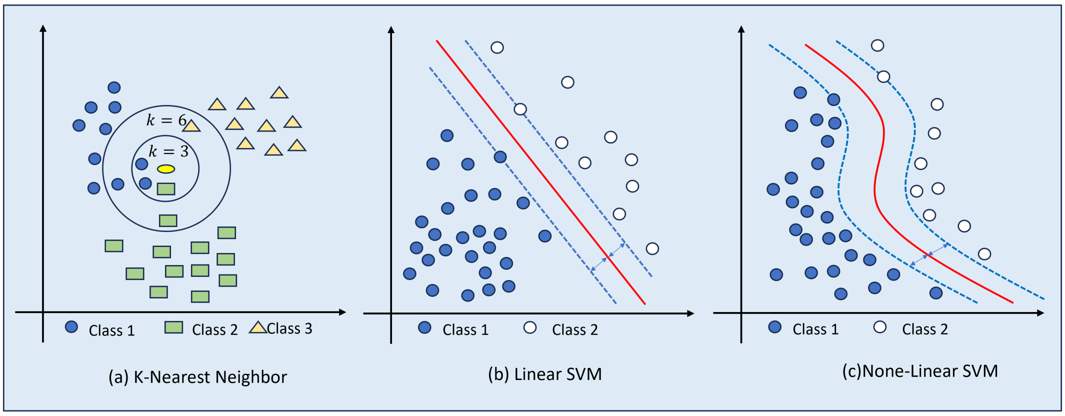


Figure 2.5: An illustration of KNN, linear SVM and none-linear SVM.

2.3.1.3 SVM-based method

Cortes and Vapnik (1994) introduced the Support Vector Machine (SVM) for binary classification in pattern recognition. Subsequently, Joachims (1998) applied SVM to text classification, which reprents the text in the form of a vector. While SVM was initially tailored for binary tasks, its application was extended to handle multi-class problems (Gou and Huang 2006). As depicted in Figure 2.5(b,c), SVM techniques transform text classification into a series of binary classification tasks, utilizing both linear and non-linear classifiers for two-dimensional datasets. In this context, the SVM creates an ideal hyperplane in a one-dimensional feature space. This hyperplane maximises the distance between two types of training sets, ensuring the model achieves optimal generalisation capability. The broader objective is to increase the distance between category boundaries in a direction orthogonal to the hyperplane, thereby minimising classification error. To ensure a globally optimal solution, The formulation of this hyperplane optimisation translates into a quadratic programming problem. For SVM to address non-linear problems effectively, the selection of an appropriate kernel function becomes crucial. Delving deeper into SVM's learning mechanics, Joachims (2001) presented a theoretical learning model. This model amalgamates statistical attributes with SVM's generalisation capabilities, providing a quantitative

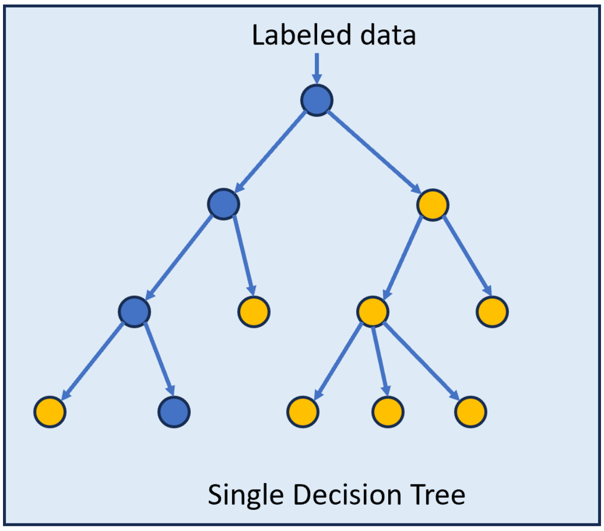


Figure 2.6: The structure of single Decision Tree.

analysis of its features and benefits. Furthermore, to minimise misclassifications in specific test collections, the Transductive Support Vector Machine (TSVM) (Joachims 1999) was introduced. This method uses prior knowledge to mould an optimal structure, thus facilitating faster learning for unique test sets.

2.3.1.4 Decision Trees

Decision Trees (DT) embody a supervised learning methodology that adopts a divide-and-conquer strategy. This method is adept at learning disjunctive expressions and is resilient to noisy text. As illustrated in Figure 2.6, the creation of decision trees encompasses two primary phases: tree construction and tree pruning. For a root node containing a set of instances with different attributes, the process begins at this node and then evaluates the data sample in sequence. The dataset is subsequently partitioned into distinct subsets based on divergent outcomes. Each subset forms a child node, and every terminal or leaf node symbolizes a distinct category. The crux of tree construction is to discern the relationship between classes and attributes, which is then leveraged to forecast the classifications of unforeseen instances. The rules formulated by the decision tree algorithm are explicit, and the tree's pruning mechanism effectively minimises the impact of noise. Nevertheless, a significant drawback of DT is its inadequacy in efficiently managing vast data volumes. For instance, the ID3 algorithm (Quinlan 1985) employs information gain for attribute selection at each node. This technique pinpoints the attribute with the maximal information gain value as the distinguishing attribute for that node. Building on Iterative Dichotomiser 3 (ID3), C4.5 (Quinlan 1992) devises a mapping from attributes to classes, facilitating the categorisation of previously unidentified entities. However, DT algorithms typically necessitate distinct training for individual datasets, resulting in efficiency challenges. Addressing this, in a symbolic rule system founded on DT (Johnson et al. 2001), texts are represented as vectors, determined by word frequencies, and rules derived from training data are employed for classifying similar datasets. To alleviate the computational burdens associated with the decision tree algorithms, the Fast Decision-Tree (FDT) technique (Vateekul and Kubat 2009) integrates a dual approach: pre-selecting attributes and concurrently training multiple decision trees on varied data subsets. For addressing imbalanced class scenarios a data-fusion method was used to combine the outcomes from these trees.

Summary Shallow learning, a variant of machine learning, primarily relies on pre-defined features crucial for prediction accuracy. Unlike deep learning models that may underperform on small datasets due to computational complexity constraints, shallow learning models often excel. Feature engineering, the process of extracting these vital features from raw text, requires substantial expertise and knowledge. Consequently, given their efficiency and adaptability to smaller datasets, there are still several researchers exploring the optimisation of shallow learning models for specialised domains.

2.3.2 Traditional Deep Learning Methods

Deep Neural Networks (DNNs) are advanced artificial neural networks designed to emulate the human brain's capabilities, enabling the automatic extraction of high-level features from data. In domains like speech recognition, image processing, and text comprehension, these networks have outperformed shallow learning models. Essential to this process is the meticulous analysis of input datasets to understand their nature, whether they are single-label, multi-label, unsupervised, or unbalanced. Word vectors are input into the DNN for training based on dataset characteristics until a certain condition is reached. The model's performance is then assessed through tasks like sentiment classification, question answering, and event prediction. The advent of deep learning in text classification began with the introduction of feed-forward and recursive neural networks, which showcased enhanced performance over their shallow learning counterparts. Modern techniques, including CNN (Kim 2013), RNN (Tai et al. 2014, Zhu et al. 2015, Lai et al. 2022, Johnson and Zhang 2016, Liu et al. 2016, Wang et al. 2017a) and attention mechanisms, are increasingly used for text classification. Researchers also adapt and modify methods like CNN and RNN, or employ hybrid and multi-task strategies, to enhance classification outcomes.

2.3.2.1 ReNN-based Method

In shallow learning models, feature design for each task is often laborious and time-intensive. On the other hand, as depicted in Figure 2.7(a), ReNN inherently grasps text semantics and syntax tree structures, eliminating the need for explicit feature crafting. With ReNN, each word in the input text acts as a leaf node of the model. Child nodes can then be integrated into the parent node through consistent weight centring. Importantly, the dimensionality remains uniform between parent and leaf nodes. The model progressively combines all nodes into one central node, representing the entire sentence. Then this representation will be used to predict the label. Correspondingly, the quantity of weight matrices corresponds to the potential children a node might possess. The representation of a parent node can be determined by computing the cumulative product of the weight matrices W_i and the representations of the children c_i, followed by the application of the transformation f.

$$h = f\left(\sum_{i=1}^{N} W_i c_i\right) \tag{2.8}$$

where N is the number of children c_i.

ReNN-based models offer enhanced performance over shallow learning models, largely because they negate the need for feature designs specific to various text classification tasks, resulting in labor cost savings. The Recursive Autoencoder (RAE) (Socher et al. 2011) is adept at predicting sentiment label distributions for given sentences, while also mastering the representations of multi-word expressions. Building upon this, the Matrix-Vector Recursive Neural Network (Socher et al. 2012) was proposed to discern representations of phrases and sentences. It's particularly versatile, handling varied lengths and types of input texts, and for each node in its constructed parse tree, it designates both a matrix and a vector. To delve deeper into capturing sentence semantics, the Recursive Neural Tensor Network (RNTN) (Socher et al. 2013) was introduced. Due to its architecture being constructed with the tree structure, RNTN processes phrases of differing lengths, representing them using parse trees and word vectors. However, the RNTN does present challenges: constructing its textual tree is time-consuming, and mapping inter-document relationships within its tree structure can be intricate. Recognising the potential for performance enhancement in deeper architectures, Irsoy and Cardie (2014) introduced the Deep Recursive Neural Network (DeepReNN). This model, constructed using binary parse trees, encompasses multiple recursive layers and is proficient in capturing the intricate nuances of linguistic compositionality.

2.3.2.2 MLP-based Method

In a technical context, a Multilayer Perceptron (MLP) is a foundational neural network architecture commonly referenced in literature as a "vanilla" neural network (Alsmadi et al. 2009). Illustrated in Figure 2.7(b) is a three-layer representation of an MLP, comprising an input layer, a single hidden layer, and an output layer. The connections between nodes are associated with specific weights. Contrary to shallow learning models, MLP processes input texts using a bag-of-words approach, often resulting in superior performance across various text classification benchmarks. MLPs are trained discriminatively using the standard back-propagation algorithm, leveraging

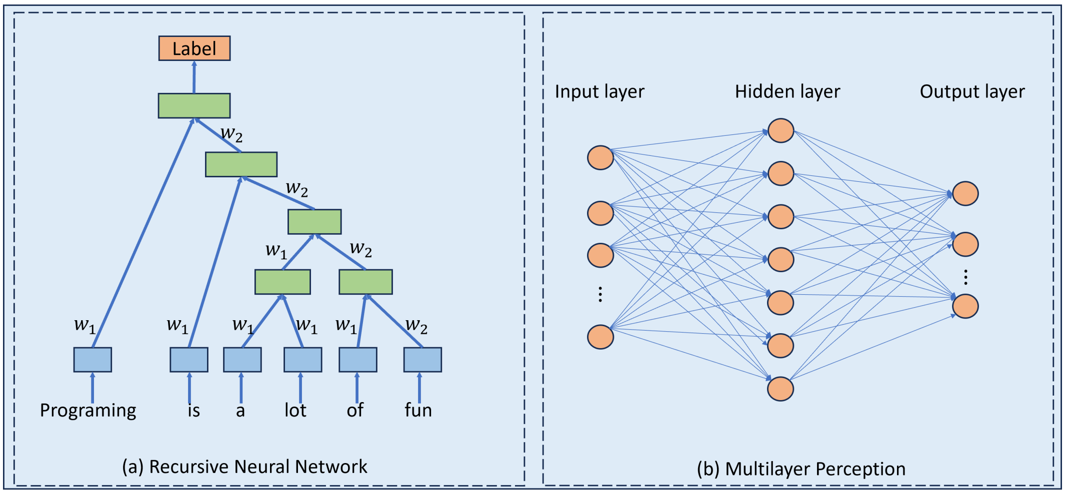


Figure 2.7: The structure of ReNN and MLP.

activation functions such as sigmoid, tanh and ReLU. For multi-class classification scenarios, the output layer typically employs a *Softmax* function. For the output vector of a hidden layer $\mathbf{x} = [x_i]_{i=1,...,n} = [x_1, x_2, ..., x_n]$, the formal definitions for each utilised activation functions are presented in the following.

The Sigmoid algorithm

$$f(x_i) = \frac{1}{1 + e^{-x_i}} \in (0, 1) \tag{2.9}$$

The ReLU algorithm

$$f(x_i) = \begin{cases} max(0, x_i), \text{ if } x_i >= 0 \\ 0\text{,if } x_i < 0 \end{cases} \tag{2.10}$$

The Tanh algorithm

$$f(x_i) = \frac{e^{x_i} - e^{-x_i}}{e^{x_i} + e^{-x_i}} \tag{2.11}$$

The Softmax algorithm

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \forall i \in (1, 2, ..., n) \tag{2.12}$$

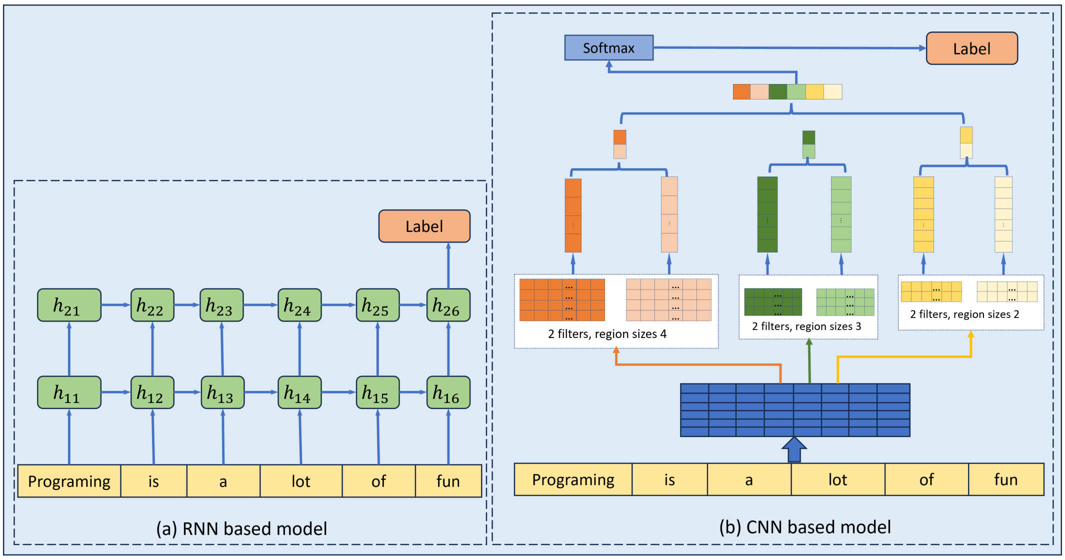


Figure 2.8: The structure of RNN and CNN.

In recent research, several MLP-based methods have been introduced into text classification tasks. Among these, the Paragraph Vector (Paragraph-Vec) (Le and Mikolov 2014) stands out as the most renowned technique, drawing parallels with the Continuous Bag of Words model (Mikolov et al. 2013a). Unlike traditional methods that struggle with varying input lengths, Paragraph-Vec leverages unsupervised algorithms to achieve fixed-length feature representations of texts. A notable distinction from Continuous Bag of Words model is its introduction of a paragraph token that is associated with a matrix-mapped paragraph vector. This model harnesses the combined or averaged vector of three word contexts to predict a fourth word. Furthermore, these paragraph vectors serve dual roles: capturing thematic memory of paragraphs and functioning as an input for prediction classifiers.

2.3.2.3 RNN-based Method

In the realm of neural networks, the Recurrent Neural Network stands out for its ability to capture long-range dependencies via recurrent computations. Particularly beneficial for text classification tasks, the RNN model is adept at assimilating historical data and factoring in the relative positioning of

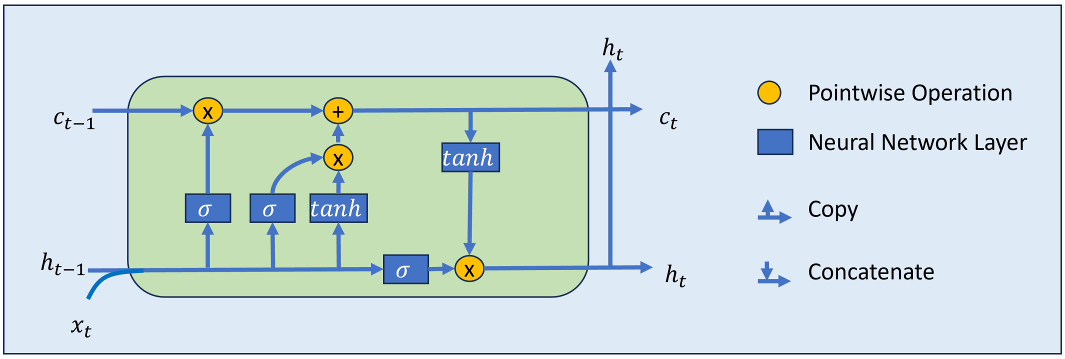


Figure 2.9: The structure of memory cell for LSTM.

words. As illustrated in Figure 2.8(a), the RNN-based text classification model functions as follows: Initially, individual words are transformed into unique vectors using word embedding techniques. These vectors are sequentially inputted into RNN cells. The outputs from these cells maintain the same dimensional consistency as their inputs and are then channelled to the subsequent hidden layer. Notably, RNN maintains uniformity in parameters across the model, ensuring each input word possesses identical weights. The final output from the hidden layer determines the predicted label for the input text.

In RNNs, the backpropagation process involves adjusting weights based on gradients derived from successive multiplications of derivatives. When these derivatives are exceedingly small, they can lead to the vanishing gradient issue. The LSTM network was introduced by Hochreiter and Schmidhuber (1997), which effectively addresses this problem. LSTMs consist of a cell designed to retain values over varied time intervals and three gates that manage the flow of information: the input, forget, and output gates. These structures allow LSTMs to adeptly discern relationships among contextual feature words. Moreover, the forget gate selectively filters out irrelevant information, enhancing the overall classification capability. Since the LSTM model has a big effect on the text tasks, therefore a comprehensive examination of LSTM's mechanics will be provided subsequently.

The compact forms of the equations for the forward pass of an LSTM cell are shown as follows:

The forgetting gate, denoted as f_t, determines which aspects of C_{t-1} contribute to C_t. Represented as a vector, each element of f_t lies between 0 and 1. Typically, the *Sigmoid* activation function is used, which produces outputs within this [0,1] range. Yet, when examining a trained LSTM, it's notable that most gate values are either close to 0 or 1, with a few in between. Among all, the operation represented by $\bigotimes$ is fundamental to LSTM's gating mechanism, illustrating the multiplicative interaction between f_t and C_{t-1}. The forgetting gate f_t can be described as:

$$f_t = \sigma(W_f \cdot [x_t, h_{t-1}] + b_f) \tag{2.13}$$

The value $\widetilde{c_t}$ signifies the updated state of the unit, derived from the input data x_t and the previous hidden node h_{t-1} using a neural network layer. Typically, the *tanh* activation function is employed for this update (Gers and Schmidhuber 2001, Gers et al. 2002). Meanwhile, i_t is termed the 'input gate'. Similar to f_t, it's a vector with elements ranging between [0,1]. This input gate is computed using x_t and h_{t-1} via the *Sigmoid* activation function.

$$i_t = \sigma(W_i \cdot [x_t, h_{t-1}] + b_i) \tag{2.14}$$

$$\widetilde{C}_t = \tanh\left(W_C \cdot [x_t, h_{t-1}] + b_C\right). \tag{2.15}$$

The update of the cell state vector are:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{2.16}$$

The final step is to determine the predicted value and prepare the full input for the next time slice by computing the output h_t from the hidden node. This output h_t is derived from the combination of the output gate o_t and the unit state C_t. The way computes o_t is similar to how it computes f_t and i_t.

$$o_t = \sigma(W_o \cdot [x_t, h_{t-1}] + b_o) \tag{2.17}$$

$$h_t = o_t \cdot \tanh{(C_t)}. \tag{2.18}$$

The Tree-LSTM model, introduced by Tai et al. (2014), evolves the conventional sequence-based LSTM architecture to accommodate tree structures. Notably, this adaptation allows subtrees with minimal impact on outcomes to be pruned via the LSTM's forgetting gate. In natural language inference, the goal is to determine if one text's meaning can be deduced from another, evaluating the semantic alignment of sentence pairs. Expanding on this, Wang et al. (2017b) introduced Bilateral Multi-perspective Matching (BiMPM) for the natural language inference task. This model first leverages a Bidirectional LSTM (BiLSTM) encoder to process input sentences. Subsequently, these encoded sentences are juxtaposed bidirectionally. A subsequent BiLSTM layer compiles these matchings into a uniform-length vector, which is then evaluated via a dense layer to derive the final outcome.

2.3.2.4 CNN-based Method

Originally intended for image classification, CNNs use convolving filters to capture image characteristics. These networks have an advantage over RNNs because they can concurrently apply various kernels, processing multiple sequence portions simultaneously. This capability has been extended to numerous natural language processing tasks, including text classification. In this context, texts are vectorized similarly to images, allowing feature extraction from multiple perspectives, as depicted in Figure 2.8(b). Initially, input tokens are represented as vectors. These vectors are then transformed into a matrix. Subsequently, this matrix is processed through convolutional layers containing filters of varying dimensions, typically between 2 to 5 dimensions. The convolutional output undergoes pooling, and the results are concatenated to generate the text's final vector representation. This vector is subsequently used to predict the text's category.

Kim (2013) introduced an unbiased model of convolutional neural networks, termed TextCNN, to apply CNNs to the text classification task. This model excels at identifying discriminative phrases in the max-pooling layer by employing a single convolution layer and keeps word vectors static, thus learning only the necessary hyperparameters. Recognising that solely relying on labelled data might be insufficient for deep models, Johnson et al. proposed a semi-supervised two-view CNN text classification method. Their approach initially utilises unlabeled data to train text region embeddings before integrating labelled data (Johnson and Zhang 2015). It is known that deep Neural Networks often deliver superior performance, but at the cost of increased computational complexity. Addressing this concern, a Deep Pyramid Convolutional Neural Network (DPCNN) is presented by Johnson and Zhang (2016). This architecture enhances computational accuracy slightly while increasing the network's depth. Notably, DPCNN distinguishes itself from ResNet (He et al. 2015) by ensuring all shortcuts are straightforward identity mappings, obviating the need for dimension matching.

In text embedding, techniques are grouped by the smallest unit: character, word, and sentence. Character-level embeddings tackle out-of-vocabulary issues, whereas word-level ones capture the syntax and meaning of words. On the other hand, sentence-level embeddings emphasise the interplay between sentences. Drawing inspiration from these studies, a character-level representation method (Nguyen and Nguyen 2017) was proposed. This approach harnesses a dictionary-based deep learning technique, integrates semantic rules, and employs deep CNNs to enhance the richness of word embeddings. Furthermore, a Transfer Capsule Network (TransCap) (Chen and Qian 2018) encapsulates sentence-level semantics into capsules, facilitating the transfer of document-level knowledge.

2.3.2.5 Attention-based Method

In the realm of text classification, both CNNs and RNNs have demonstrated outstanding performance. Nevertheless, a notable drawback of these architectures is their limited interpretability, often making it challenging to elucidate classification errors due to the obscurity of hidden layers. To address this, attention-based mechanisms have emerged as a prominent solution. Initially introduced by Bahdanau et al. (2014) for machine translation, the attention mechanism's potential was later harnessed by Yang et al. (2015) to develop the Hierarchical Attention Network (HAN). HAN incorporates two encoders paired with dual layers of attention. This innovative design allows the model to selectively focus on pertinent input segments. The architecture first consolidates crucial words into sentence representations and subsequently synthesises these into comprehensive text vectors. Drawing from real clinical conversations between doctors and patients, symptoms significantly influence the final decision. Therefore, when processing medical free text, symptom-related information should be given more importance than ordinary words. The attention mechanism addresses this prioritisation. Through this dual-level attention, the model effectively discerns the significance of each word and sentence, thereby enhancing interpretability and applicability in classification tasks.

The attention mechanism has garnered popularity in text classification due to its ability to enhance performance while providing interpretability. Various research works have built upon this mechanism. For instance, LSTMN (Cheng et al. 2015) processes text sequentially from left to right, leveraging memory and attention for rudimentary reasoning. Similarly, an attention-based LSTM neural network was proposed (Wang et al. 2015), exploring the interplay between input sentences and their respective aspects. Zhou et al. (2015) proposed the BI-Attention for crosslingual text classification, designed to identify bilingual long-range dependencies. Additionally, Hu et al. (2018) put forth an attention mechanism that is tailored based on category attributes, specifically addressing the challenge of imbalances in datasets with few-shot charges.

2.3.3 Transformer-based language models

The Transformer architecture (Vaswani et al. 2017), revolutionised NLP tasks by providing a deep learning model adept at processing sequences. Unlike traditional RNNs and LSTMs that process data sequentially, transformers employ the attention mechanism to capture global dependencies between inputs and outputs, enabling parallel processing. This advancement has significantly improved performance in various NLP tasks, including text clas-

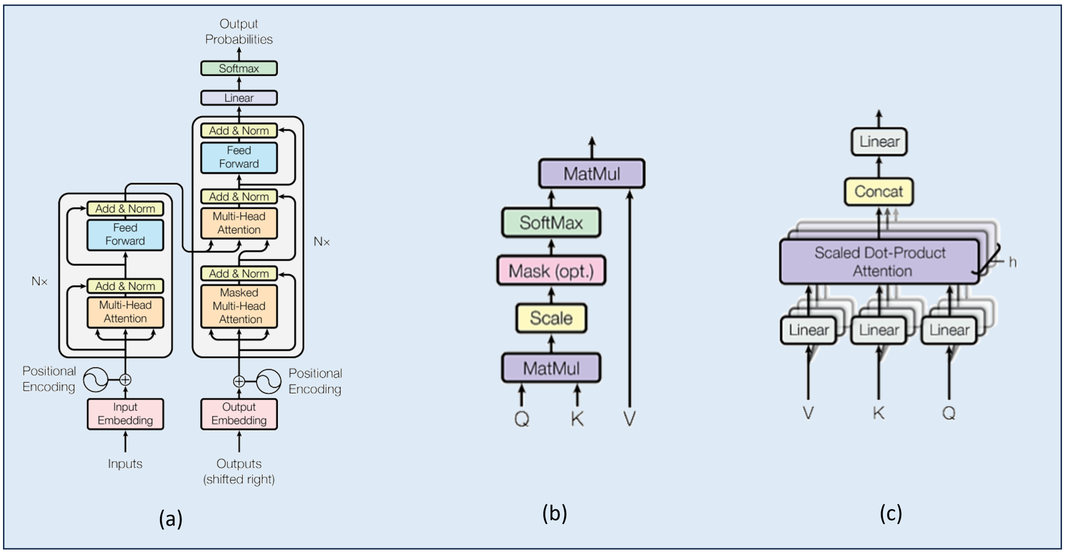


Figure 2.10: Illustration of Transformer. (a) The architecture of Transformer; (b) Scaled Dot-Product Attention; (c) Multi-Head Attention **(Vaswani et al. 2017)**

sification. Structurally, a transformer model is bifurcated into an encoder and a decoder, each composed of stacked identical layers. As illustrated in Figure 2.10(a), both the encoder and decoder modules integrate a multi-head self-attention mechanism. Additionally, the encoder features a position-wise feed-forward network and an “Add & Norm” residual connection. The decoder, while mirroring these features, also incorporates multi-head attention over the encoder’s output. Positional encoding is embedded at the base of both modules, ensuring the model recognises token sequences, a critical addition since the self-attention mechanism is inherently permutation-invariant. Since the methods used in this thesis are transformer-based, the subsequent sections will delve deeper into the specifics of the transformer’s architecture.

Multi-head Self-Attention

In the Transformer model, rather than relying on a single set of attention weights, multiple sets are employed to enable the model to concurrently attend to various segments of the input, catering to different tasks or motivations. Through the application of several self-attention mechanisms in tandem, distinct output matrices are generated. These matrices are subsequently merged and undergo a linear transformation. The core component of this process is the multi-head self-attention structure, depicted in Figure 2.10(c). Within this mechanism, given a query Q, a key K, and a value V, the attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.19}$$

Denoted by h, the number of these projection heads, the attention mechanism is further parameterised by d_k, which represents the depth of the queries and keys. The output of multi-head is presented as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{2.20}$$

where each head is:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.21}$$

here, W_i^Q, W_i^K, and W_i^V are parameter matrices.

Position-wise Feed-Forward Networks

In the encoder and decoder architecture, each layer incorporates a fully connected feed-forward network alongside attention sub-layers. For each position, this network is applied identically and independently. It is structured as two consecutive linear transformations, separated by a ReLU activation function, as shown in the following:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{2.22}$$

where x is the output of Add & Norm layer, W_1, W_2, b_1 and b_2 are parameters for two consecutive linear transformations. In the proposed technique, linear transformations are consistently applied across various positions, but they differentiate in parameters between layers. This can equivalently be characterised as two 1x1 convolutions. Both the input and output have a dimensionality of $d_{model} = 512$, whereas the internal layer boasts a dimensionality of $d_{ff} = 2048$.

Positional Encoding

In the model, the inherent lack of awareness regarding the token sequence necessitates the addition of a positional encoding to the input embeddings. Specifically, for a given position *pos* and dimension i, the positional encoding is incorporated to address this order insensitivity.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right) \tag{2.23}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{2.24}$$

Add & Norm

In the model, each sub-layer, whether it's self-attention or feed-forward, incorporates a residual connection that is subsequently followed by layer normalisation. When given an input x, the output from any sub-layer is denoted as $SubLayer(x)$.

$$LayerOut(x) = \text{Norm}(x + \text{SubLayer}(x)) \tag{2.25}$$

where $Norm$ is a method that stabilises the output from each layer, ensuring a consistent average and range. For a input z, Here's how it functions for a particular feature vector:

$$\text{Norm}(x) = \frac{z - \mu}{\sigma} \times \gamma + \beta \tag{2.26}$$

In this context, μ represents the mean of z, while σ is its standard deviation. The learnable parameters, γ and β, provide the model with added flexibility. This enables the model to adjust and potentially restore the original scale and positioning of the data.

Final Linear and Softmax

In the decoder's output, a final linear layer processes it. Then, using x as input, a softmax function produces probabilities for the target vocabulary, as shown in the following:

$$P(x) = \text{softmax}(xW_x) \tag{2.27}$$

where W_x is a learned weight matrix.

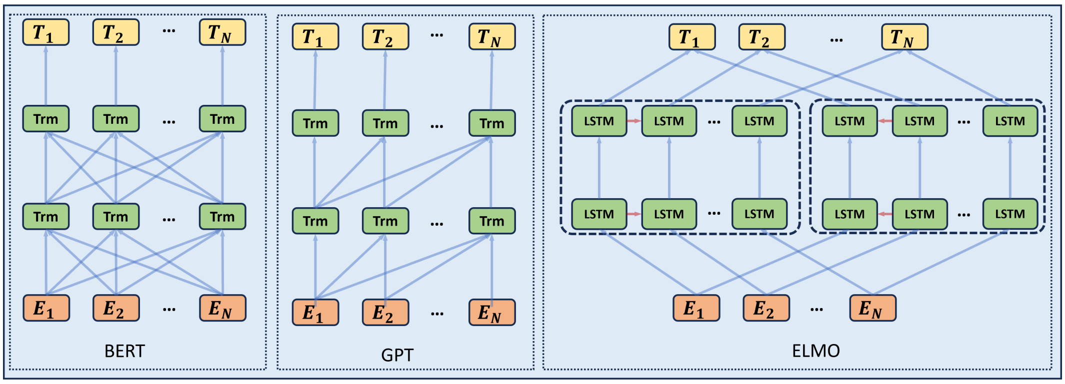


Figure 2.11: Structure of the Transformer-based models **(Devlin et al. 2018a)**, including BERT, OpenAI GPT and ELMo. The token embedding vectors of the input are represented by E_i. 'Trm' stands for the transformer block, 'LSTM' refers to the LSTM block, and T_i is the predicted target.

2.3.4 Pre-trained language models

In the realm of NLP, the advent of BERT (Devlin et al. 2018b) represented a significant milestone. This model's ability to produce contextualised word vectors has been instrumental in enhancing text classification techniques, among other NLP tasks. Subsequent studies have shown that BERT-based text classification models consistently outperform their predecessors across various tasks. Figure 2.11 presents model architectures of BERT (Devlin et al. 2018a), GPT (Radford et al. 2018) and ELMo (Peters et al. 2017). They will be discussed in the sections that follow.

ELMo, known for its deep contextualised word representations, readily integrates into various models. Its strength lies in capturing intricate word characteristics by providing unique embeddings for diverse linguistic settings. This is achieved using a bi-directional LSTM to contextualise each word based on surrounding words.

GPT utilises both supervised fine-tuning and unsupervised pre-training to garner general representations. These representations are adept at adapting to numerous NLP tasks with minimal customisation, even if the target task's domain diverges from the unlabeled datasets. Typically, GPT's training comprises two phases. The first entails deriving initial neural network model parameters from an unlabeled dataset. Subsequently, these parameters are refined using a supervised objective tailored for the specific target task. Recently, the successful implementation of ChatGPT and GPT-4 (OpenAI 2023) in real customer scenarios propels the zero/one/few-shot learning of LLMs into an entirely new phase. Notable examples include GPT-3 with approximately 175B parameters (Brown et al. 2020), PaLM with around 540B parameters (Chowdhery et al. 2022), Galactica with close to 120B parameters (Taylor et al. 2022), and LLaMA, which ranges from 7B to 65B parameters (Touvron et al. 2023). These LLMs are fundamentally based on the Transformer architecture (Vaswani et al. 2017), incorporating deep neural networks with stacked multi-head attention layers.

BERT, a model pioneered by Google, emphasizes pre-training deep bidirectional representations from unlabeled text. This is accomplished by considering both the left and right context in each layer. Besides, using the Masked LM task and Next Sentence Prediction (NSP) task for training gives the model a strong ability for generalisation. As a result, BERT has enhanced various NLP task performances, especially in areas like text classification. The fine-tuning process in BERT involves appending a single output layer, thereby facilitating the development of models for a range of downstream NLP tasks, including sentiment analysis, question-answering, and machine translation.

In essence, ELMo uses an LSTM-based feature strategy, whereas both BERT and GPT rely on fine-tuning approaches with the Transformer architecture. Notably, ELMo and BERT adopt bidirectional training, whereas OpenAI GPT follows a left-to-right training paradigm. Owing to its synthesis of ELMo and OpenAI GPT's strengths, BERT often achieves superior outcomes.

In recent years, transformer-based models have gained significant traction in the domain of NLP owing to their ability to parallelize computation, making them apt for handling large-scale datasets. These models do not solely rely on sequential information, making them a preferred choice for NLP tasks.

Among these transformer architectures, RoBERTa (Liu et al. 2019b) stands out with its dynamic masking technique, where a new masking pattern is generated for every sequence before feeding it into the model. This approach is complemented by a more extended pre-training phase, where RoBERTa also assesses the impact of diverse hyperparameters and the volume of training data. ALBERT (Lan et al. 2019), on the other hand, employs two-parameter simplification techniques. It is noteworthy that many of these transformer-based methods utilise unsupervised objective functions during the pretraining phase. These functions, such as the next sentence prediction, masking, and permutation, have proven to be potent in understanding word dependencies and semantic structures (Jawahar et al. 2018). XLnet (Yang et al. 2019a) brings a novel perspective with its generalised autoregressive pre-training method. It aims to maximise the expected likelihood across all permutations of factorisation orders, ensuring effective learning of bidirectional contexts. What's more, it addresses certain limitations of BERT through its autoregressive formulation and successfully incorporates concepts from Transformer-XL (Dai et al. 2018) during its pre-training phase. Lastly, in the realm of domain-specific language representation models, BioBERT (Lee et al. 2020) emerges as a leader. Primarily pre-trained on expansive biomedical corpora, BioBERT retains a consistent architecture across different tasks. The results are commendable: it not only outperforms BERT but also establishes new benchmarks by surpassing other leading models in numerous biomedical text mining challenges.

2.3.5 Others.

QA style for the sentiment classification task: Sentiment classification can intriguingly be approached as a Question-Answering (QA) task. In this context, Shen et al. (2017) presented a novel methodology supported by a meticulously annotated corpus. In the hierarchical matching network, they presented, three stages effectively capture the details between questions and answers.

Prompt-based classification : Prompt-based classification leverages the generalisation capabilities of LLMs, enabling them to perform diverse classification tasks without task-specific training. However, effective utilisation requires careful prompt design, and sometimes, fine-tuning on specific tasks to achieve optimal performance. This approach has been effectively used in numerous NLP applications, offering a versatile and data-efficient methodology for text classification. The implementation of prompt-based methodologies within the medical field primarily aims to predict diagnostic outcomes, categorise conditions, and assign relevant tags (Lamichhane 2023, Sivarajkumar and Wang 2022, Wang et al. 2023a, Elfrink et al. 2023, Zhang and Chen 2023, Akrout et al. 2023, Zhu et al. 2023, Chen et al. 2023, Lin et al. 2023). Utilising a streamlined approach, these methods facilitate an efficient and precise identification and categorisation of diseases or conditions, thereby supporting healthcare professionals in rendering informed decisions and optimising patient care strategies. Consequently, the straightforward application of such approaches paves the way for a more accessible, structured, and systematic exploration and management of healthcare data.

2.4 Evaluation Metrics

Typically, in text classification tasks, accuracy and F1 score are the primary metrics used to evaluate methods. They are presented as follows:

Accuracy is a metric used in classification to determine the proportion of correctly predicted observations to the total observations. It is a measure that gives you an overview of how well your classifier is performing. However, accuracy is not always the best metric to rely on, especially when dealing with imbalanced datasets.

$$Accuracy = \frac{TP + TN}{N} \tag{2.28}$$

where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative.

F1 score is a measure of a test's accuracy that considers both precision and recall. It is particularly useful when the class distribution is unbalanced.

The F1 Score is the harmonic mean of precision and recall, where an F1 Score reaches its best value at 1 (perfect precision and recall) and the worst at 0. To understand the $F1$ score, precision and recall need to be known:
Precision (also called positive predictive value) is the number of true positives divided by the number of all positive predictions (including both true positives and false positives).

$$Precision = \frac{TP}{TP + FT} \tag{2.29}$$

Recall (also known as sensitivity or true positive rate) is the number of true positives divided by the number of all actual positives (including both true positives and false negatives).

$$Recall = \frac{TP}{TP + FN} \tag{2.30}$$

Then, the F1 score is given by:

$$F1 = \frac{2 * Precision \times Recall}{Precision + Recall} \tag{2.31}$$

It essentially captures a balance between how accurately the model predicts the positive class and how good it is at capturing actual positive instances.

2.5 Summary

This section delves into various models for text classification, spanning data pre-processing methods, shallow learning, deep learning, and transformer-based methodologies. We begin by presenting the process of data processing and then introduce the primary models from the realm of shallow learning, highlighting their essence in enhancing text classification through refined feature extraction and classifier design. Subsequently, a transition is made to explore the paradigms of traditional deep learning models and contemporary transformer-based architectures. Both of these deep learning approaches encompass multiple hidden neural network layers, exemplifying a high degree of intricacy. Their distinct advantage is their capability to discern feature representations directly from input data, eliminating the need for extensive manual intervention or reliance on prior knowledge. However, a salient characteristic of these models is their data-driven nature, which necessitates vast datasets to deliver optimal performance. While transformer-based models, underpinned by self-attention mechanisms, provide a semblance of interpretability by capturing relations among words, their explanatory power falls short when juxtaposed against the intuitive nature of shallow models.

Chapter 3

Knowledge-based dialogue system for self-diagnosis

3.1 Background

The healthcare system today faces significant challenges. As artificial intelligence and big data learning become more prevalent, their applications in medical diagnosis offer potential relief. To mitigate these pressures on the health system, two primary strategies emerge. First, automatic diagnosis tools for patients can offload some demands from the medical system by delivering initial diagnostic services. Second, offering diagnostic support tools for physicians can enhance their efficiency and expertise, further elevating the overall effectiveness of the medical system. Collaborating with the doctor at Nuffield Health Bournemouth Hospital, this chapter delves into the pipelines of a diagnostic dialogue system, examining feasible solutions for each module in terms of medical data acquisition, data processing, knowledge graph building, diagnosis policy design, and dialogue system construction.

In order to obtain the symptom information of patients and better prevent diseases in advance, many self-diagnosis systems were proposed. According to the survey (Semigran et al. 2015), almost 35% of U.S. adults regularly use self-diagnosis tools (i.e., Symptomate[1], Patient[2], HealthTap[3], Ada[4] and

[1]https://symptomate.com/
[2]https://patient.info/symptom-checker
[3]https://www.healthtap.com/
[4]https://ada.com/

etc.) on the internet. Furthermore, a 2016 report on Internet usage[5] for health-related self-diagnosis among adults in the United Kingdom showed that approximately 41% of individuals had used online platforms for self-diagnosing health issues within a short period.

That is mainly because of its lower cost and wide accessibility, which makes a growing number of adults self-diagnose their illness or disease through automatic diagnosis tools at the first instance instead of walking directly into a hospital. It also makes automatic diagnosis more and more popular.

Most automatic medical diagnosis systems are designed to complete the process of patient consultation in the form of an online web or mobile app. Such a system is easy to access, and patients do not need to go to the hospital, which saves both time and financial cost. Besides, the automatic diagnosis system can give reasonable treatment suggestions to patients with common diseases, which relieves the pressure on the health service.

Automatic diagnosis is a very challenging topic in the field of digital health. The previous studies on this topic mainly focused on Electronic Health Records (EHR), while dialogues between doctors and patients that contain more rich information were not well studied. As described by Lewenberg et al. (2017), in a typical real-patient clinical consultation process, the patient first describes his or her feelings and symptoms as a self-report to initiate the conversation. Based on the patient's self-report, the doctor extracts explicit symptoms and asks about other implicit symptoms, using his or her professional medical knowledge to make a diagnosis. This is a multi-step reasoning process. At each step, the doctor chooses a symptom to ask about or concludes the diagnosis by considering the dialogue history and possible diseases. The logic components of real patient checking are shown in Figure 3.1.

There are three methods to complete an automatic diagnosis process, the online search-based method, the ontology-based method and the dialogue-based diagnosis method.

[5]https://www.statista.com/statistics/605362/average-frequency-of-internet-self-diagnosis-uk/

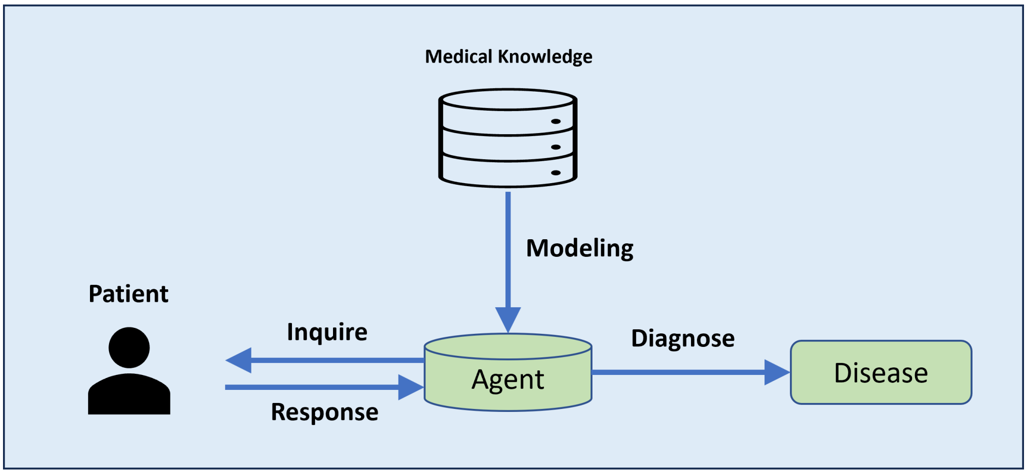


Figure 3.1: The logic components of real patient checking (Tang et al. 2016)

As for the **online search-based method**, the self-diagnosis process is referred to as symptom checking. It allows patients to search on the internet with search engines and find the solution for their situation. Because of the "no-certified" information on the internet, online search-based services for self-diagnosis often return irrelevant information and sometimes absurd results, which will bring great risks to patients.

Most of the **ontology-based method** is mainly constructed with decision trees, the symptom-disease graph or rule-based expert systems, which provide several options for patients to match their symptoms and guide them to complete the self-diagnosis step by step. Although this type of system has a higher diagnostic accuracy, its shortcomings are still obvious. High accuracy requires the system to understand more details about the patient, and all possible symptoms will be considered regardless of their possibilities. Therefore, based on the work of Lewenberg et al. (2017), Shim et al. (2018), a large number of questions from the system would be asked to obtain the symptom values, which is inefficient and time-consuming for the patient.

The conversation between a patient and a doctor can provide many valuable clues for the diagnosing of the patient's symptoms. During the process of conversation, the doctor can quickly diagnose the patient's disease based on his/her own experience. Besides, Bickmore and Giorgino (2006) proposed that the one-on-one, face-to-face interaction between doctor and patient is widely acknowledged to be the "gold standard" for the process of patient consultation. Therefore, the dialogue system is naturally used in the field of automatic medical diagnosis to simulate the conversation between the doctor and patient. This type of dialogue system (Kao et al. 2018) has great potential to improve the efficiency of collecting information from patients and assist general practitioners in clinical diagnosis.

A dialogue system, also known as an interactive conversational agent, virtual agent or sometimes chatbot, is a computer system intended to converse with a human. It is widely used in technical support services, language learning tools and entertainment. With the massive growth of internet data, the fast improvement of hardware computing capabilities, and the continuous breakthrough of NLP and deep learning technologies, dialogue systems have boomed during the last few years. Dialogue systems, especially task-oriented dialogue systems can not only support users with recommended information (e.g., searching restaurants (Wen et al. 2016) and booking tickets (Li et al. 2017, Peng et al. 2018)) and complete specific tasks (e.g., voice control of loT devices) but also can interact with users like human beings. Because of its potential commercial value, it has attracted increasing attention and has been used in different domains.

In recent years, with the improvement of the task-oriented dialogue system, it was also applied to the field of intelligent medical diagnosis. Wei et al. (2018) cast the dialogue systems as Markov Decision Process (MDP), and firstly proposed the task-oriented medical dialogue system to determine diseases using Reinforcement Learning (RL) based method. The **dialogue-based diagnosis system methods** heavily rely on the complex belief tracker (which is used to estimate the patient's goal at every step of the dialogue) and data-driven learning, which cannot be applied to automatic diagnosis directly because of insufficient data for training. Xu et al. (2019) proposed a knowledge-routed relational dialogue system for automatic diagnosis, which not only uses Deep Q-network (DQN) via data-driven learning to manage topic transitions (deciding which symptoms should be asked) but also takes the relations among diseases and symptoms into consideration.

However, the medical knowledge graph from Xu et al. (2019) heavily relies on statistical features (i.e., conditional probabilities from symptoms to diseases). The scale of their training data is limited (only 423 conversation data are selected as the training set), which makes it hard to provide key insights into symptom-disease relations and still insufficient in the utilisation of medical knowledge.

The ontology-based diagnosis system primarily employs decision trees, symptom-disease graphs, or rule-based expert systems, emphasising the relationships between symptoms and diseases. While its accuracy is commendable, the system's efficiency remains a challenge. Dialogue-based diagnostic methods, through patient interaction, can predict forthcoming symptoms and diagnose more efficiently. This kind of system needs a lot of real conversational datasets for training the model, however, the available conversational datasets for training these dialogue systems are limited and primarily sourced from online consultations, affecting their reliability and thus limiting their functionality.

Consequently, two primary issues need to be addressed for the automatic diagnostic dialogue system. Firstly, due to the limitation of the conversational dataset, a credible knowledge-based dataset or one derived from actual clinical patient scenarios should be established. Secondly, a decision-making strategy must be introduced to enhance consultation accuracy and efficiency.

In the context of the Intel-PA project, which is detailed in Chapter 1, my role entailed translating the pragmatic requirements provided by the doctors into an academic framework and subsequently proposing academically oriented solutions. Therefore, the aim of this chapter is to investigate pipelines of an efficient diagnostic dialogue system, examining feasible solutions for each module in terms of medical data acquisition, data processing, knowledge graph building, diagnosis policy designing and dialogue system construction.
Contribution This study contributes to the field of automatic diagnosis by offering the following insights:

- Investigation of the knowledge-based automatic diagnosis dialogue system pipelines with proposed solutions for each module, the demonstration video can be found at `https://github.com/ruibin-wang/Dialogue-based-self-diagnosis.git`.

- Presentation of a flowchart to construct a disease-symptoms graph from textual descriptions of symptoms.
- Provision of an open-source disease-symptoms dataset derived from the official NHS website, available at the github link above.
- Development of a two-option dialogue policy to enhance diagnostic efficiency.

3.2 Related work

3.2.1 Benchmark datasets

The lack of training data is a common problem of the dialogue system, not to mention the dialogue data between doctors and patients in the consultation process. In order to advance the research in the medical dialogue system for automatic diagnosis, building a dataset which contains the annotated conversation between doctors and patients in the consultation process is very necessary. Since Wei et al. (2018) first proposed the task-oriented dialogue system for automatic diagnosis in 2018, there are currently 5 public medical dialogue datasets as far as we know, including MuZhi proposed by Wei et al. (2018), DX constructed by Xu et al. (2019), CMDD launched by Lin et al. (2019), MIE built by Zhang et al. (2020a) and MedDG proposed by Liu et al. (2020a). The details of these datasets are noted in Table 3.1.

The doctor-patient conversation data in these 5 datasets is collected from the health community and forums like Chunyu-Doctor[6] and MuZhi-Doctor[7] where patients can describe their symptoms and then have a short online text-based dialogue with qualified doctors for professional diagnosis. The collection of these benchmark datasets has been completed and well organised for the automatic diagnosis system. To some extent, these databases alleviate

[6] https://www.chunyuyisheng.com/
[7] https://muzhi.baidu.com/

Table 3.1: The details of 5 current public dialogue diagnosis datasets

Dataset	**Domain**	**Number of diseases**	**Number of dialogues**	**Number of Symptoms**
Muzhi	Pediatrics	4	710	70
DX	Pediatrics	5	527	46
CMDD	Pediatrics	4	2067	161
MIE	Cardiology	6	1120	71
MedDG	Gastroenterology	12	17864	160

the predicament of data scarcity, but for intelligent diagnosis systems based on dialogue, these are far from enough.

Direct dialogues between patients and doctors are infrequently published, primarily due to the challenges in recording and the expenses associated with anonymization. This scarcity poses challenges in training dialogue systems. While a significant portion of existing data is sourced from online consultations, its credibility can be questionable. In contrast, official knowledge-based datasets detailing disease-symptom relationships are more reliable.

Two primary knowledge-based datasets outline the relationship between diseases and symptoms: the clinical knowledge graph[8] and the Kaggle disease-symptoms dataset, including "Diseases and Symptoms[9]", "Disease Symptom Prediction[10]" and "Disease Symptoms and Patient Profile Dataset[11]". The clinical knowledge graph offers an extensive understanding of diseases, but its vast size often includes descriptive symptom details, making it challenging to align with specific patient scenarios. On the other hand, the Kaggle disease-symptoms dataset, detailed in Table 3.2, covers a narrower range of diseases and their associated symptoms. As a result, there is a pressing need

[8]https://ckg.readthedocs.io/en/latest/INTRO.html

[9]https://www.kaggle.com/datasets/pjmathematician/diseases-and-symptoms?resource=download

[10]https://github.com/itachi9604/healthcare-chatbot/blob/master/Data

[11]https://www.kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset

to develop a comprehensive dataset that is both credible and inclusive of prevalent diseases and their symptoms.

Table 3.2: The collected disease-symptoms datasets from Kaggle

Name	**Description**	**Downloadable**
Diseases and Symptoms	134 diseases with 408 symptoms	Yes
Disease Symptom Prediction	41 diseases with 132 symptoms	Yes
Disease Symptoms and Patient Profile Dataset	116 diseases, 10 symptoms	Yes

3.2.2 Bayesian Inference and Tree-Based Methods

There has been a lot of work in the past involving self-diagnosis. One large family of models uses Bayesian inference and tree-based methods (Kononenko 1993 2001, Xu et al. 2013, Kohavi et al. 1996), which utilise entropy functions to pick symptoms based on the information gain theory. Although these models are able to handle medical diagnosis tasks, the feature which can be used to perform a prediction is hard to collect and its acquisition requires very heavy computation from processing the data. Therefore, in order to reduce the cost of feature acquisition, Nan et al. (2015) introduced the decision tree and feature-budgeted random forest methods into this field. His method can minimise prediction error for a user-specified average feature acquisition budget. In addition, Hayashi (1990) tried to extract rule-based representations from medical data and human knowledge for the purpose of performing medical diagnosis. Since the global maximisation of the information gain is intractable, these methods often use greedy algorithms or approximations that result in low predictive accuracy.

3.2.3 Task-oriented dialogue-based diagnosis systems

Task-oriented systems (Glass et al. 1995, Seneff et al. 1998, Rudnicky et al. 1999, Levin et al. 2000, Walker et al. 2001 2002, Raux 2005, Andreani et al.

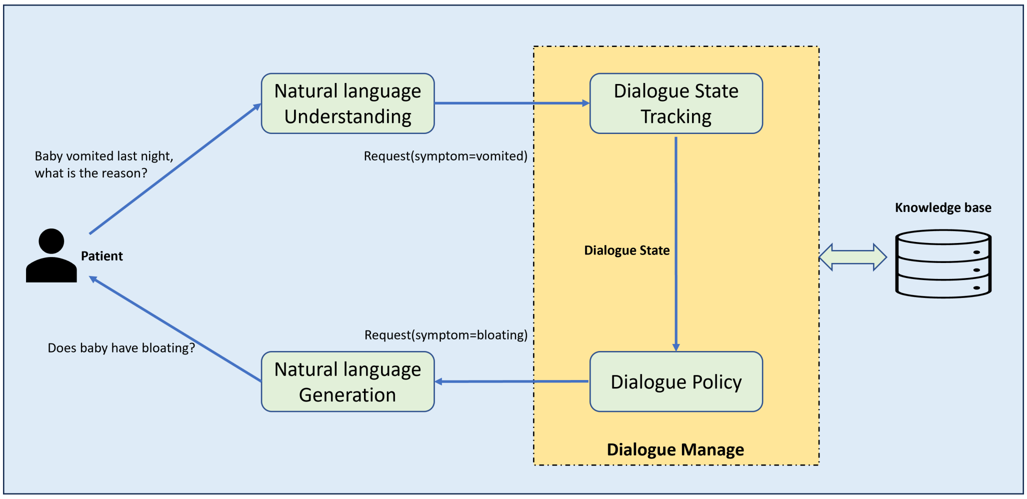


Figure 3.2: Pipeline of task-oriented dialogue system

2006, Wang et al. 2011) often have a specific goal to achieve, and it is designed to assist users in solving a specific task such as ranging the restaurant reservation, scheduling the meeting, planning the business and completing the special instructions. These systems are usually constrained to specific domains. Existing research on task-oriented dialogue systems can be roughly divided into two categories: pipeline and end-to-end methods (Zhang et al. 2020b).

As illustrated in Figure 3.2, the pipeline method consists of a Natural Language Understanding (NLU) module, a Dialogue Management (DM) module (including Dialogue State Tracking (DST) and Dialogue Policy) and a Natural Language Generation (NLG) module. Before training the dialogue policy component and aggregating NLU, DST and NLG components together, they are usually trained separately.

The details of these modules are described below:

- Natural Language Understanding (NLU): this component is designed to map the user's input utterance to a structured semantic representation which is decomposed into two tasks, including the user's intent (for example, asking for suggestions about disease diagnosis) detection and slot-value (i.e. name of symptoms) extraction. The former research mainly views intent detection as a classification task and the slot-value task as a sequence labelling problem:

$$p_{intent} = (d|x_1, x_2, ..., x_n) \tag{3.1}$$

$$p_{s-v} = (y_1, y_2, ..., y_n|x_1, x_2, ..., x_n) \tag{3.2}$$

where the d indicates the class of intent, x_i is the token in the utterance $[x_1, x_2, \ldots, x_n]$ and the corresponding y_i is its label. Due to the good performance of sequence modelling, several neural network-based methods are proposed in the NLU task, including RNN and its variants (Yao et al. 2013 2014, Hakkani-Tür et al. 2016), CNN (Xu and Sarikaya 2013) and recursive neural network (Guo et al. 2014). These methods use the hidden state of x_i to predict the corresponding label y_i and utilise the final hidden state of the whole utterance $[x_1, x_2, \ldots, x_n]$ to classify the user's intent d.

- Dialogue Management (DM): The DM module is used to estimate the user's goal and generate the next system action in each step by taking the entire dialogue context as input. Generally, the DM module is constructed with DST and Policy, where the DST section is utilised to estimate the state of the current dialogue based on a given history of dialogues and the Policy section is designed to generate the next system action. Since the dialogue acts of a task-oriented system are generated sequentially, the whole process is always viewed as a Markov Decision Process. Early work set the dialogue state as a fixed set, which makes the system difficult to expend. In order to solve this problem and make the system more robust in sophisticated situations, Young et al. (2013), Young (2006), Schatzmann et al. (2007) proposed a Partially Observable Markov Decision Processes (POMDPs) framework. With the development of machine learning and neural networks, Reinforcement Learning (RL) (Li et al. 2016b, Shi et al. 2019) is also used to train the dialogue policy of the system, which shows a good performance in training the dialogue policy.

- Natual Language Generation (NLG): The function of this component is to map the system action to a natural language utterance as a response to the user's inquiry. It is often modelled as a conditioned language generation task (Wen et al. 2015). Peng et al. (2020) first proposed a Semantically-Conditioned Generative Pre-Training (SC-GPT) method based on the pre-training GPT with large-scale NLG corpus to make the generated utterance more natural and informative. For the purpose of reducing the complexity of the system, most automatic diagnosis systems use the template-based method for the NLG module. Generally, the syntax of the system responses for this template-based method is fixed, and only a specific position can fill the system actions.

Although the modular structure of the pipeline system makes it more interpretable and stable, each part of the pipeline system is trained separately and then combined to complete the task, which makes errors in this system gradually accumulate. In order to improve the accuracy of the pipeline system, more recent work focuses on building an end-to-end system where each component is joined in a large neural network (Sarikaya et al. 2016, Wen et al. 2016). However, the structure of the end-to-end system is a black box, which is more uncontrollable.

In recent years, with the improvement of the task-oriented dialogue system, it has also been applied to the field of automatic medical diagnosis. Based on the framework of a task-oriented dialogue system, Kao et al. (2018) regarded the patient consultation process as a sequential decision problem and utilised an RL-based hierarchical Deep Q-learning (DQN) method to solve it. Besides, Wei et al. (2018) built a medical diagnosis dialogue dataset that contains 67 symptoms and detects four types of diseases. Considering the relationship between diseases and symptoms, Xu et al. (2019) proposed a Knowledge-routed Deep Q-learning Network (KR-DQN) for medical diagnosis that reduces the inquires of illogical and repeated symptoms. However, these works are data-driven methods and highly rely on statistical features, which is still insufficient for medical diagnosis because of data scarcity. Unlike the above models, without usable conversation data for training the dialogue system, this chapter proposes a knowledge-based dialogue system for automatic diagnosis.

3.3 Dataset

In the absence of conversation data, the focus of this chapter is to explore the pipelines for dialogue-based diagnosis systems. A potential solution involves the collection of a knowledge-based disease-symptoms dataset and the construction of a rule-based system. A flowchart has been developed to outline the process of constructing a disease-symptoms graph from descriptive symptom texts. This flowchart is divided into two main sections: Data Processing and Symptom Extraction and Disease-Symptom Graph Construction, detailed in subsequent sections.

3.3.1 Data Processing and Symptom Extraction

The NHS of the United Kingdom provides a dedicated website offering guidance on the conditions, symptoms, and treatments of common diseases. The data from this NHS website[12] is extracted to construct the disease-symptoms graph. As the guidance is presented in a free-text format, the data requires processing to yield symptoms suitable for diagnosis. An illustration of the web extraction and symptom identification is provided in Figure 3.3.

In Figure 3.4, a flowchart depicts the process from data collection to the construction of a disease-symptoms graph. Initially, using Python-based tools and frameworks such as “urllib”, “beautiful soup (bs4)”, “Scrapy”, and “Crawley”, the raw symptom descriptions for specific diseases is obtained. Subsequently, symptom extraction tools are employed to extract relevant symptom data from these descriptions. Notably, various methods, including BIO, BioBERT (Lee et al. 2020), MedCAT (Kraljevic et al. 2021), John Snow Clinical Annotator, IBM Watson Clinical Annotation, and others, are

[12]https://www.nhs.uk/conditions/

Figure 3.3: Illustration of web scrawl (left) and symptom extraction (right)

available for extracting medical entities from sentences. While some high-performing methods come at a cost, others that are freely available might not offer optimal performance. Among these, IBM Watson has made significant advancements in medical AI and provides a complimentary API known for its excellent extraction performance. Hence, the IBM Watson Clinical Annotator API was chosen for symptom extraction. The final step involves using the extracted disease and symptom data to construct the disease-symptoms graph within the graph store database.

3.3.2 Disease-Symptom Graph Construction

After reviewing the NHS guidelines for common diseases and extracting symptoms using the Watson API, the data on **268 diseases and their 807 associated symptoms** is obtained. The disease-symptom pairs are structured as a graph. Given that Neo4j facilitates rapid and efficient querying of linked data, surpassing the speed of traditional relational databases, it is chosen for data storage. Figure 3.5 displays the disease-symptom pairs for “Myelodysplastic syndrome”, “Keratosis pilaris”, “Bedbugs”, and “Polycystic ovary syndrome”. In this figure, diseases are represented in yellow circles, while their associated symptoms are shown in blue circles.

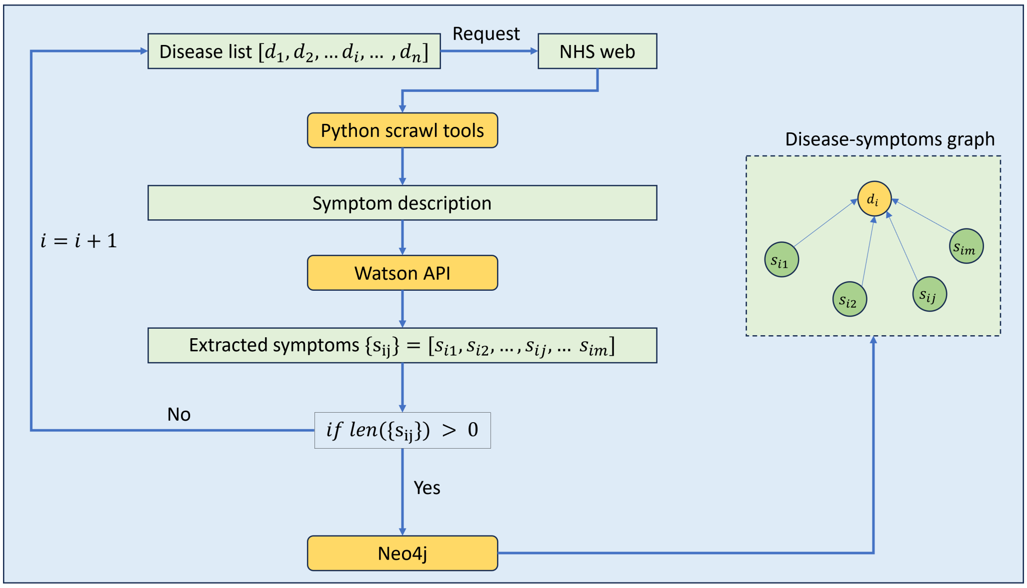


Figure 3.4: Flowchart of data processing and disease-symptoms graph building

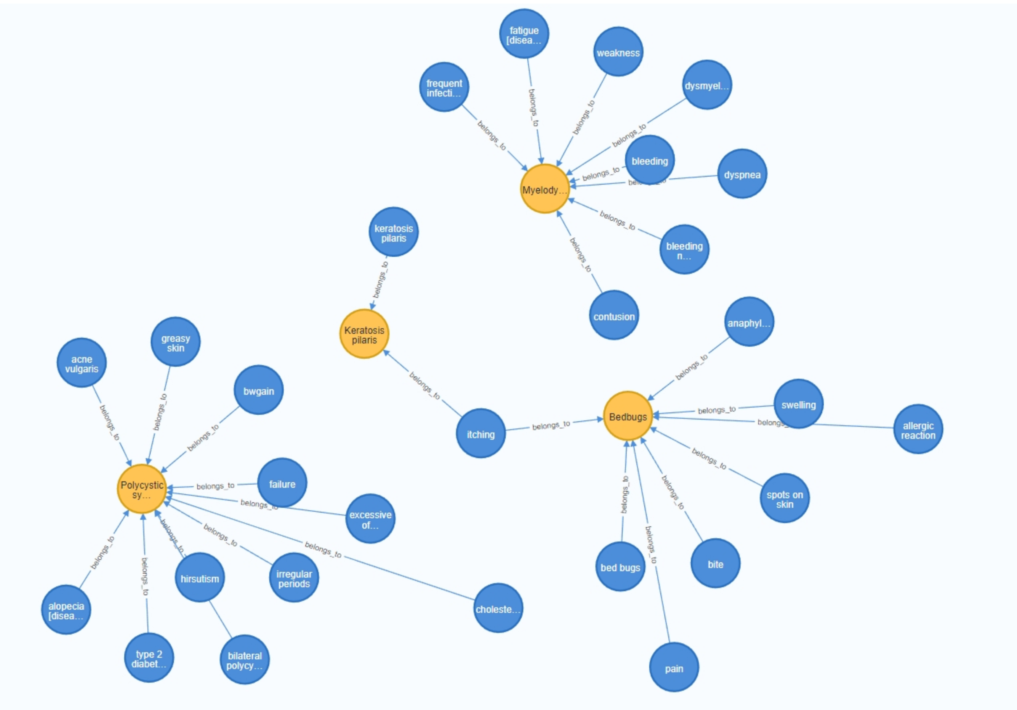


Figure 3.5: The illustration of disease-symptoms graph in the Neo4j

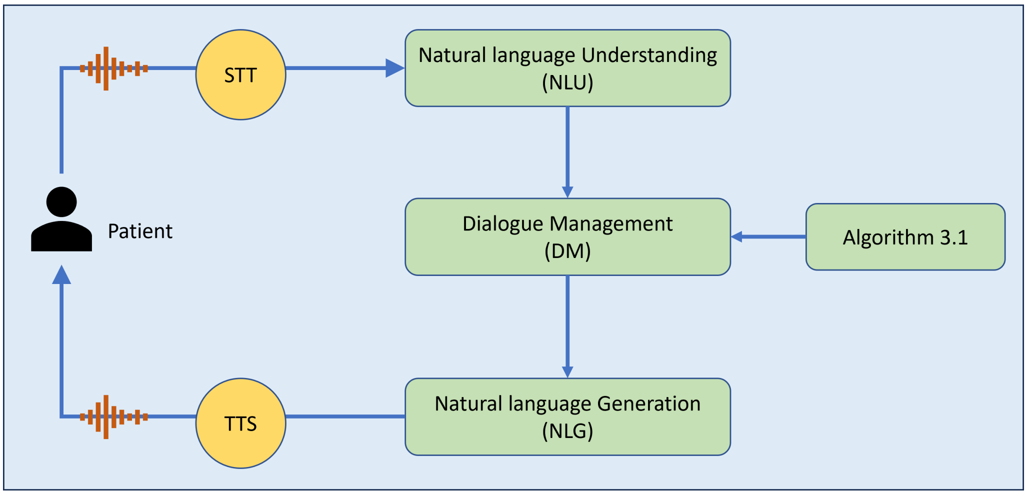


Figure 3.6: Framework of the knowledge-based dialogue system for self-diagnosis

3.4 System overview

In the "Related Work" section, a dialogue system framework was introduced. Building upon that architecture, Figure 3.6 provides an overview of the knowledge-based dialogue system for disease diagnosis. The proposed system distinguishes itself by incorporating two additional modules: Speech-to-Text (STT) and Text-to-Speech (TTS). The designs facilitate bidirectional conversion between voice and text, enhancing convenience for patients. Moreover, the dialogue management module operates under the two-choice diagnosis policy, as detailed in Algorithm 1. The subsequent sections will elucidate the functionality of these modules.

3.4.1 STT and TTS

To reduce the complexity of the proposed dialogue-based diagnosis system, the state-of-the-art Google Cloud API is used to convert patients' speech to text. Besides, although the speeches generated by online TTS are smoother and more natural, the delay caused by data transmission will reduce the efficiency of system action. Therefore, to improve the real-time response of the proposed dialogue-based diagnosis system, offline Google Text-to-Speech (gTTs) was used in the proposed system.

3.4.2 NLU

The module identifies patient intent and extracts symptoms from their descriptions. In the "Data collection" section, several methods such as BIO, BioBERT (Lee et al. 2020), MedCAT (Kraljevic et al. 2021), John Snow Clinical Annotator, and IBM Watson Clinical Annotation are discussed for medical entity extraction from sentences. IBM Watson was chosen for this module due to its superior performance in clinical entity extraction.

3.4.3 Dialogue Management

In this module, using the acquired disease-symptoms graph, a rule-based strategy was devised for automated diagnosis through patient dialogue. As the dataset is based on fixed knowledge without real cases for training, a traditional decision-making method is introduced for disease diagnosis. A brief overview and definition of this algorithm are provided below, with comprehensive details presented in Algorithm 1.
The conversation starts with:

- **Turn 0:** The system provides a response, prompting the patient to describe their situation.

- **Processing:** Upon receiving the patient's "Description", symptoms within this description are extracted using the following equation:

$$known\_symptoms = extract\_symtoms(Description) \tag{3.3}$$

  Here, *extract_symptoms* refers to clinical entity extraction tools. The IBM Watson clinical annotation tool, known for its extensive research and state-of-the-art performance in medical text processing, was employed for this purpose. An example of this process can be seen as follows:

Description: We would be grateful if you could review this 56-year-old gentleman. He was admitted on 27-11-18 after falling down the stairs. Preceding dizziness where he felt he would pass out but did not lose consciousness. 2 weeks ago, states developed sudden onset headache - vice-like, severity at start 8-10. Not positional. Right eyes have a tunnel vision. On examination had reduced abduction right eye and diplopia. The case was discussed with Dr xxx who suggested MRI. The MRI report is as below and suggests a neurology review.

known_symptoms: [falling down, dizziness, pass out, headache, tunnel vision, diplopia]

The system then identifies associated disease nodes on the graph based on the symptoms in *known_symptoms*. This can be represented as:

$$D = [D_1, D_2, \ldots, D_i, \ldots, D_k], i = 1, 2, \ldots, k \tag{3.4}$$

Here, k denotes the number of diseases related to the extracted symptoms. Each D_i in D encompasses related symptoms, expressed as:

$$D_i = [s_1^{D_i}, s_2^{D_i}, \ldots, s_{m_{D_i}}^{D_i}] \tag{3.5}$$

Where m_{D_i} indicates the number of symptoms associated with the disease D_i.

To enhance the efficiency of the consultation process by providing an extensive range of symptom options for patients, a score function was established for each disease, represented as D_i^{score}. The score may be seen as a probability of disease susceptibility; a higher score correlates with an elevated likelihood of disease contraction. Furthermore, a sorted symptoms set, S^{sort}, was introduced, defined as follows:

$$D_i^{\text{score}} = \frac{n_{D_i}^{\text{know\_symp}}}{\text{len}(D_i)} = \frac{n_{D_i}^{\text{know\_symp}}}{m_{D_i}} \tag{3.6}$$

$$S^{\text{sort}} = [s_1^{\text{sort}}, s_2^{\text{sort}}, \ldots, s_n^{\text{sort}}] \tag{3.7}$$

Here, $n_{D_i}^{know\_symp}$ denotes the count of symptoms present in both D_i and *known_symptoms*. The set S^{sort} encompasses all symptom nodes in the disease-symptoms graph, where n indicates the total number of symptoms. Symptoms are organised by their degree in the graph, such that:

$$degree(s_1^{sort}) > degree(s_2^{sort}) > \ldots > degree(s_n^{sort}) \tag{3.8}$$

Given that the disease nodes have an associated score, they can be arranged based on this score, leading to a new definition:

$$D^{sort} = [(D_1, D_1^{score}), (D_2, D_2^{score}), \ldots, (D_i, D_i^{score}), \ldots, (D_k, D_k^{score})] \tag{3.9}$$

Where:

$$D_1^{score} > D_2^{score} > \ldots > D_i^{score} > \ldots > D_k^{score} \tag{3.10}$$

- **Turn 1 to Turn N:**

  The system presents patients with a binary choice, for instance, "Do you have s_1 or s_2?" where s_1 and s_2 represent symptoms identified from the constructed disease-symptom graph. The strategies employed to select these symptoms aim to facilitate an efficient preliminary diagnosis. Specifically, s_1 is selected from D_1 because D_1 comprises the most symptoms from the *known_symptoms*. Conversely, s_2 is derived from S^{sort}, a dataset populated with prevalent symptoms ranked by their degree; hence, in many cases, there's a heightened likelihood for a patient to manifest this symptom. For clarity, an reorder function has been established that adjusts the ranking of symptoms or diseases based on their scores in descending order:

$$list = reorder(list) \tag{3.11}$$

The comprehensive decision-making process is delineated in Algorithm 1. The consultation procedure will conclude when either of the following conditions is met:

$$turn = N \text{ or } D_i^{score} = 100\% \tag{3.12}$$

where N is the maximum turn of dialogue.

Algorithm 1 The two-choices diagnosis policy

Definition: $S^{sort} = [s_1^{sort}, s_2^{sort}, \ldots, s_n^{sort}]$ is the sorted symptom nodes set, *extract_symtoms* in the symptom extraction tool, D_i^{score} is the score of disease D_i, D^{sort} is the connected disease nodes ordered by D_i^{score}.

turn 0:

$\overleftarrow{system}$: *What can I do for you? Please give a description of your situation.*

$\overrightarrow{user}$: *Description*

Processing:

$known\_symptoms = extract\_symptoms(Description)$

for *symptom* in *known_symptoms* **do**

  Extract all related diseases, set them as new D^{sort}

end for

for *symptom* in *known_symptoms* **do**

  remove *symptom* from D_i, $D_i = reorder(D_i)$

  remove *symptom* from S^{sort}, $S^{sort} = reorder(S^{sort})$

end for

turn 1 to turn N:

for *turn* in $range(1, N)$ **do**

  $s_1 = s_1^{D_i} \in D_1$

  remove $s_1^{D_i}$ from S^{sort}, $S^{sort} = reorder(S^{sort})$

$s_2 = s_1^{sort} \in S^{sort}$, $s_2 \in D_i$

$\overleftarrow{system}$: *Do you have* s_1 *or* s_2*?*

$\overrightarrow{user}$: *Response*

if *both* in *Response* **then**

  $known\_symptoms.append(s_1, s_2)$

else if *first* in *Response* **then**

  $known\_symptoms.append(s_1)$

else if *last* in *Response* **then**

  $known\_symptoms.append(s_2)$

else if *none* in *Response* **then**

  pass

end if

$D_1^{score} = reorder(D_1^{score})$, $D_i^{score} = reorder(D_i^{score})$

remove $s_1^{D_1}$ from D_1, $D_1 = reorder(D_1)$

remove s_1^{sort} from D_i, $D_i = reorder(D_i)$

remove s_1^{sort} from S^{sort}, $S^{sort} = reorder(S^{sort})$

$D^{sort} = reorder(D^{sort})$

if $D_i^{score} = 100\%$ or $turn = N$ **then**

  $\overleftarrow{system}$: *You might have* D_1*, with a possibility of* D_1^{score}*,* D_2 *with a possibility of* D_2^{score}*.*

end if

end for

3.4.4 NLG

In this module, a template-based method is presented for patient interaction. This approach offers clear instructions to the patient while eliminating extraneous information. The system response for this method can be found in Algorithm 1.

3.5 Experiment and Results

3.5.1 Demo demonstration of Algorithm 1

Using Algorithm 1, the demo was enhanced to a web application. This interface facilitates interaction between the patient and the system, supporting both text and voice as input and output formats. A screenshot of the knowledge-based dialogue system is provided in Figure 3.7. The content of this diagnosis process is displayed as follows:

- Dialogue content in Figure 3.7

  [System]: What can I do for you? Please give a description of your situation.

  [Patient]: Hello, I'm not feeling well and am currently sick. I have a history of chronic kidney disease, and today, I'm experiencing diarrhoea. It seems that I may be dehydrated as well.

  [System]: Do you have feeling queasy or pain?

  [Patient]: both

  [Patient]: Do you have confusion or swelling?

  [Patient]: none

  [System]: Do you have confusion or elevated core body temperature?

  [Patient]: first

  [System]: Do you have drowsy or itching?

  [Patient]: first

  [System]: Do you have illness or headache?

  [Patient]: both

  [System]: You might have
  Kidney injury: with a possibility of 0.8

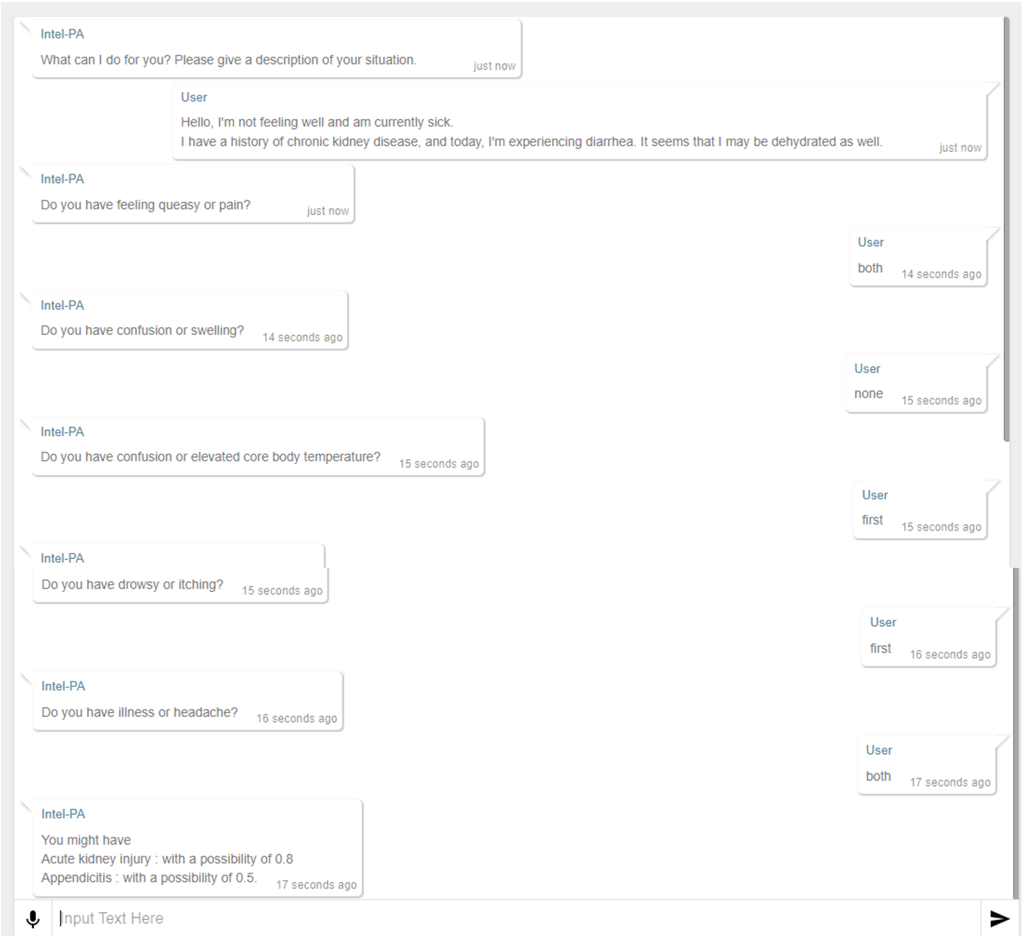


Figure 3.7: Example of self-diagnosis process of two-choice policy-based dialogue system

Appendicitis: with a possibility of 0.5.

The present demonstration selects "Kidney Injury" as the focal point for diagnosis, given its associated symptoms, which include *acute kidney failure, feeling queasy, sick, diarrhoea, dehydration, confusion, drowsy, end-stage renal disease (ESRD), illness, chronic kidney disease.* A demonstration video is available at the following link: `https://github.com/ruibin-wang/Dialogue-based-self-diagnosis.git`.

3.6 Summary

In this chapter, the construction of a dialogue-based diagnostic system for patients is examined, encompassing stages such as data collection, processing, graph building, and the design of both diagnosis policies and the dialogue system. Disease description data were sourced from the NHS official website to construct a disease-symptoms graph. The graph presents 268 diseases and their corresponding 807 symptoms, exceeding the volume of the collected dataset. This representation clarifies the associations between diseases and their symptoms and is archived in a Neo4j database for optimised retrieval and querying. Leveraging the disease-symptoms graph, a knowledge-based two-choice diagnostic policy was developed for the system, with subsequent web app demonstration examples provided. It is worth mentioning that owing to the development schedule and the intricate policy of the collaborated hospital, the evaluation of the developed web applications in this chapter—in genuine clinical contexts is presently under development by my colleagues.

Chapter 4

Hybrid Architecture Based Intelligent Diagnosis Assistant for GP

Note that this chapter has been previously published under the same title in the Journal "BMC Medical Informatics and Decision Making," with a DOI reference: `https://doi.org/10.1186/s12911-023-02398-8`.

4.1 Background

In many healthcare systems, including the United Kingdom and Ireland, GPs are the first point of contact for patients and most general problems are addressed in primary care (O'Donnell 2000). GP plays a central role in ensuring that the patient receives a timely and preliminary diagnosis. However, if GP is unable to make decisions, the patient will be referred to a specialist.

For a patient eligible for referral, GP will write a referral letter to a specialist, outlining the reason why this patient is being referred as well as the patient's past medical history, including medications, allergies, and any other relevant information. After 2 weeks (urgent situations) or an an 18-week (non-urgent) wait, an appointment with a hospital specialist will be offered to the patient.

With the outbreak of COVID-19, the NHS is under extreme pressure. Hospital Episode Statistics (HES) data shows that the number of unsuccessful GP referrals has jumped from 238,859 in February 2020 to a staggering 401,115 in November 2021 (an 87% increase), because there is no capacity in secondary care, and those referrals are rejected. During the specialist consultation process, almost two-thirds of the time is spent on confirming the contents of the referral letter and doing basic examinations, then on the professional tests and making a decision. If this initial confirmation and basic examination process are completed before the specialist consultation, this process will be accelerated which can save a lot of time and more patients will have a chance to get timely health care. GP is the most suitable role to take over this work. As the first point of contact for patients, GP has the best understanding of the patient's situation. Because GP has a broader, less specialised knowledge base, the accuracy of their primary diagnosis is limited. If there is an assistant for the GP to improve the accuracy of their primary diagnosis, it will be helpful to relieve the pressure on the specialist. A lot of research has been devoted to developing AI-assisted diagnosis methods (Lewenberg et al. 2017, Shim et al. 2018, Wei et al. 2018, Xu et al. 2019) or tools (i.e., Symptomate, Patient, HealthTap, Ada, etc.). However, they are mainly facing the patient, and the connection between the GP and specialists is seldom considered.

Therefore, the purpose of the paper is to develop an intelligent GP assistant which can improve the accuracy of the GP's primary diagnosis. The referral letter includes abundant information about the patient's past medical history, including medications, allergies, and any other relevant information. The proposed system takes advantage of the complaints in the referral letter and the final diagnosis from specialists to train an AI model. Collaborated with neurologists at the University Hospital Dorset (UHD) to collect the source datasets. The proposed architecture supports the diagnosis of four common neurological diseases, including epilepsy-recurrent seizures, headaches, dorsalgia, and cerebral infarction.

Diagnosing on the basis of a referral letter can be seen as a text classification task. With the development of research in natural language processing, there are mainly two types of models that can be used to complete this task: 1) the traditional deep learning-based text classification models including Convolutional Neural Network (CNN) (Kim 2013, Conneau et al. 2016), Long Short-Term Memory networks (LSTM) (Mousa and Schuller 2017) and their variants; 2) the pre-trained encoder-based models including Transfomer (Vaswani et al. 2017), Bidirectional Encoder Representations from Transformers (BERT) (Alaparthi and Mishra 2020), DistilBERT (Sanh et al. 2019), XLNet (Yang et al. 2019b) and Robustly Optimised BERT approach (Roberta) (Liu et al. 2019a).

Distinct from the typical text classification task, the symptoms present in the content for classification will play a pivotal role in determining the final outcome. The collected dataset not only contains the long complaint text but also includes short symptom phrases (extracted symptoms). For the traditional deep learning-based model, the feature of each word uses static (fixed) word vectors that come from the pre-trained word embedding models; they have better performance in short text classification. As embedding models themselves, the pre-trained encoder-based structures have the advantage of extracting the long-distance dependencies in the text by using a multi-head attention approach to calculate the similarities between words. Therefore, fusing them together can not only enrich the representation of the neurology dataset but also benefit from both short-text symptoms and long-text complaints.

A hybrid architecture is proposed in this chapter, which fuses the features of words in different representation spaces to make the primary diagnosis from a referral letter. This structure jointly optimises the pre-trained encoder-based model and traditional deep learning model, which makes full use of the two different representation spaces. Besides, two data augmentation methods (Complaint-Symptoms Integration Method and Symptom Dot Separating Method) have also been proposed for this architecture. Subsequent experiments demonstrate that this hybrid architecture has better performance in terms of the accuracy of the text classification task. Finally, an AI diagnosis assistant web application which leverages the superior performance of this architecture to help GPs complete referral tasks efficiently and accurately is developed. The code is available at `https://github.com/ruibin-wang/referral-letter-classification`.

The contribution of this chapter can be summarised as follows:

- This chapter proposed a **Hybrid Architecture** which fuses the features of words in different representation spaces.
- This chapter proposed a **Symptom Dot Separating Method** to avoid semantic confusion of the extracted symptoms.
- A **Complaint-Symptoms Integration Method** is proposed, which is proven to have a positive effect on the accuracy of the disease classification tasks.
- An AI diagnosis assistant **web application** is developed targeting to help GPs complete primary diagnosis efficiently and accurately.

4.2 Related work

The referral letters, encompassing detailed records of patients' historical illnesses and current symptomatic presentations, offer a valuable resource for assisting GPs in decision-making. This study introduces a novel approach, utilising referral letters as a predictive tool for disease diagnosis. To our knowledge, this is the first initiative to employ referral letters for disease diagnosis. Currently, the primary methodologies for aiding diagnosis include task-oriented dialogue systems and text classification methods. This proposed use of referral letters aims to complement these existing techniques, potentially enhancing the diagnostic process.

4.2.1 The task-oriented dialogue systems

In the realm of task-oriented methodologies, as referenced in studies (Lewenberg et al. 2017, Shim et al. 2018, Wei et al. 2018, Xu et al. 2019), the primary focus lies on developing a coherent dialogue policy. This policy aims to elicit detailed information from patients through conversation to facilitate accurate diagnoses. The foundational data in these methods is derived from dialogues between patients and doctors. It is important to note that the nature of this source data differs from the focus of the current research problem.

4.2.2 Text classification methods

The complaint in the referral letter presents the patient's medical information in a long text. Text classification methods can be used for mining useful information in referral letters and classifying it into the correct disease categories. There are three main text classification methods: machine learning-based methods, traditional deep learning-based methods and pre-trained attention-based methods.

Machine learning-based methods: Machine learning approaches in this domain primarily relied on predefined features, such as associated symptoms and disease status, as inputs. These methods employed classifiers like Support Vector Machine (SVM) (Joachims 1998), tree-based methods (Kononenko 1993 2001, Xu et al. 2013, Kohavi et al. 1996, Nan et al. 2015, Hayashi 1990), Logistic Regression (LR) (Genkin et al. 2007), and Naïve Bayes (NB) (McCallum et al. 1998) for classification purposes. In these methods, their application in the medical text classification field includes, SVM methods used for detecting adverse drug reactions (Sarker and Gonzalez 2015), identifying cancer-related causes (Koopman et al. 2015) and detecting positive cancer admissions (Kocbek et al. 2016); tree-based methods used for supporting COVID-19 (Chrimes et al. 2023) and for pregnancy diagnosis (Bahani et al. 2021); LR for diagnosing chronic kidney (Alshebly and Ahmed 2019) and hepatolenticular degeneration (Richards et al. 1996); and NB for predicting cardiovascular risk (Bandyopadhyay et al. 2015), predicting unstable angina (Vila-Francés et al. 2013), warfarin therapy management (Yet et al. 2013) and heart failure decompensation (Sarkar and Koehler 2012). However, these approaches, deeply rooted in statistical theory, often necessitate extensive feature engineering. Such reliance on manually crafted features could overlook scenarios not encompassed in the predefined sets. This limitation renders these traditional methods less effective for the classification of referral letters, where the range of features might not be counted.

Traditional deep learning-based methods: The advent of word embedding techniques, as referenced in (Joachims 1998, McCallum et al. 1998, Lin and Hovy 2003, Zhang et al. 2008, Mikolov et al. 2013b), has significantly enhanced the efficacy of traditional deep learning-based methods in text classification tasks. These methods, including Convolutional Neural Networks (CNN) (Kim 2013, Conneau et al. 2016), Long Short-Term Memory Networks (LSTM) (Mousa and Schuller 2017), and their various adaptations (Conneau et al. 2016, Zhang et al. 2015, Kim et al. 2018), leverage word embeddings to encode textual semantics into a word representation space. The application of these methods in processing medical text includes: CNN used for sentence-level classification tasks in the Merck Manual dataset (Hughes et al. 2017) and classifying nursing-care texts (Nii et al. 2017); and LSTM used for multi-label binary text classification (Oleynik et al. 2019) and clinical record assignment (Venkataraman et al. 2020). In these approaches, each word in the text is represented by static word vectors derived from pre-trained word embedding models such as fastText (Kononenko 2001), word2vec (Xu et al. 2013, Kohavi et al. 1996, Nan et al. 2015), and GloVe (Hayashi 1990). The unique filter design of CNN and the attention mechanisms within LSTM make these models particularly adept at identifying local and positionally invariant features in texts, such as key phrases or terms crucial for classification. Despite this, the architectural constraints of CNN and LSTM, coupled with the use of static word embeddings, limit their effectiveness in classifying longer texts. This limitation arises from the inherent inability of these models to dynamically adapt word representations based on the surrounding textual context, which is often essential for understanding more extended text passages.

Fine-tuning on the pre-trained attention-based methods: Word representation varies across different semantic spaces, a concept exemplified by pre-trained attention-based models like Transfomer (Vaswani et al. 2017), BERT (Alaparthi and Mishra 2020), DistilBERT (Sanh et al. 2019), XLNet (Yang et al. 2019b), Roberta (Liu et al. 2019a). The application of these methods in medical text processing includes: social media health-related text classification (Guo et al. 2022), classifying patient's medical conditions from their discharge summaries (Lu et al. 2022), and PubMed abstracts and MIMIC-II clinical notes classification Peng et al. (2019). These embedding models, trained on extensive datasets, offer rich contextual representations of words. Additionally, the incorporation of multi-head self-attention mechanisms within their architecture enables them to effectively handle long texts with complex dependencies. This capability is crucial for accurate text classification, as it allows for a more nuanced understanding of the text's semantic structure.

This research focuses on a dataset of referral letters, where the key symptoms detailed within the text are crucial for classification outcomes. Additionally, different from the plain text classification in other domains, the dependency attributes of these symptoms play a significant role in determining the final classification results. Therefore, it is essential to propose a method that can simultaneously consider both the overall context and the local key information within the text. This research presents two solutions from data augmentation and classification architecture aspects. Regarding data augmentation, the symptoms mentioned in the complaint text are extracted, and then they are consolidated together as neural network input. This process employs the Symptom Dot Separating Method and the Complaint-Symptoms Integration Method, which are thoroughly described in Section 4.3.1 on Data Pre-processing. In the above discussion on text classification methodologies, CNN and LSTM techniques excel at recognising local and positionally invariant features in text. However, their efficacy diminishes when dealing with extended text sequences. In contrast, pre-trained attention-based models show superior performance with longer texts but lack the precision to consistently highlight all key symptoms in context. Therefore, this chapter propose a hybrid architecture that involves the integration of these two model types. This hybrid architecture leverages the strengths of each model type, enhancing the system's capacity for accurate text interpretation and classification. Complemented by appropriate data augmentation methods designed to emphasize key symptoms within the referral letters, the resultant architecture benefits from the synergistic combination, leading to a more robust and nuanced performance in text analysis. The subsequent section will detail the proposed data augmentation methods and the hybrid architecture.

4.3 Methods

Given the constrained dimensions of the acquired dataset, this chapter introduces two data augmentation strategies and a hybrid architecture aimed at bolstering the vector representation of medical text. Subsequent sections will delineate the methodology, offering a clear explication of the proposed approaches.

4.3.1 Data Pre-processing

This chapter collected a dataset of referral letters from the neurologists at UHD. The collected dataset includes records of four common neurological diseases: epilepsy-recurrent seizures (G40), headache (R51), dorsalgia (M54), and cerebral infarction (I63), along with their corresponding complaint texts. An example of this dataset can be found in Table 4.2, and the dataset's statistics are presented in Table 4.1. In Table 4.2 and Table 4.1, the referral letter presents the patient's clinical condition and medical history information in the complaint text. The codes G40, R51, I63, and M54 correspond to the International Classification of Disease (ICD) codes. As shown in Figure 4.1, before training the hybrid architecture, the complaint text in the referral letter is processed with the following four steps:

Table 4.1: Statistic of the neurology dataset.

Disease code	**Number of referral letters**
G40	79
R51	65
M54	59
I63	58
Total	261

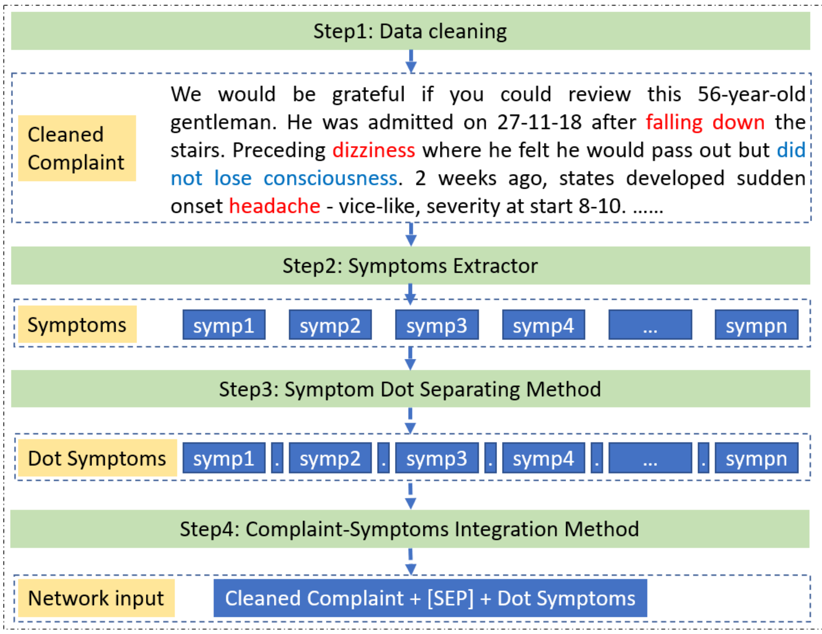


Figure 4.1: Flowchart of the data pre-processing

Step1: Data cleaning

Upon analysis of the collected referral letters, this chapter identified extraneous characters and phrases, including salutations and symbols like $$, **, @, ??, and XXX. Consequently, this section implemented a step to detect and remove this irrelevant information.

Step2: Extract symptoms

Symptoms are the key information for a GP to make decisions. A clear and abundant description of symptoms will improve the accuracy of diagnosis. Therefore, this step is to extract symptoms in the given text. Although MIE (Zhang et al. 2020a), BERN (Lee et al. 2020) and Stanza (Zhang et al. 2021) can recognize some symptoms, the accuracy is not satisfactory, and negated symptoms cannot be recognized. This chapter adopts the Watson Annotator for Clinical Data of IBM (note that this tool has been unavailable since the start of January 2023 due to service adjustments), which can provide a high recognition accuracy for symptoms. Negated symptoms are symptoms that the patient does not have, but which have been mentioned in the complaint text, they will be removed in this step, for example, the blue text in Table 4.2.

Table 4.2: Example of complaint text processing.

Process	**Text**
Complaint text	We would be grateful if you could review this 56-year-old gentleman. He was admitted on 27-11-18 after falling down the stairs. Preceding dizziness where he felt he would pass out but did not lose consciousness. 2 weeks ago, states developed sudden onset headache - vice-like, severity at start 8-10. Not positional. Right eyes have a tunnel vision. On examination had reduced abduction right eye and diplopia. The case was discussed with Dr xxx who suggested MRI. The MRI report is as below and suggests a neurology review.
Processed symptoms	falling down. dizziness. pass out. headache. tunnel vision. diplopia
Diagnosis	R51, Headache
Network input	Complaint + [SEP] + processed symptoms

Step3: Symptom Dot Separating Method

Most symptoms consist of multi-words such as falling down, pass out and tunnel vision in Table 4.2. If these symptoms are connected directly in a sequence as the input for models, the relation between connection words (for example the word 'down' and 'dizziness' in the processed symptoms part of Table 4.2) will be calculated, which will extract confusing information from the sequence and has a negative effect on the accuracy of disease classification. Therefore, this chapter proposed a **Symptom Dot Separating Method**, using an unmeaningful symbol dot ('.') to integrate the extracted symptoms into training, validating and testing data. Lateral experiments will prove that under the same network structure, the classification performance of training data processed with the Symptom Dot Separating Method is better than that without using this method. The processed symptoms can be seen in Table 4.2.

Step4: Complaint-Symptoms Integration Method

For medical text classification, there are two approaches to training models. The first approach takes advantage of the high performance of pre-trained models in processing long text by directly embedding (Zhou et al. 2021) the complaint text. The second approach firstly extracts the key information (symptoms) in the complaint text and then uses the extracted symptoms as input to train the model (Shim et al. 2018, Wei et al. 2018, Xu et al. 2019).

The first training method treats each word equally, which loses the focus on key information in the complaint text. The second training method utilises the extracted symptoms, abandoning other valuable information in the sentence. Therefore, this chapter proposed a **Complaint- Symptoms Integration Method**, which combines the complaint text and processed symptoms with a simple symbol "[SEP]". This symbol is often used in pre-trained encoder-based models as a special connection between two different sentences. The proposed integration method not only keeps the original semantics of the patient's historical medical information but also highlights key information in the complaint text. Subsequent experiments have shown that the classification accuracy of this method is higher than that of training complaint text and extracted symptoms alone.

Table 4.2 shows an example of complaint text processing. It can be clearly seen that the related symptoms are correctly extracted (marked in red in the complaint text), while negated symptoms are also obtained (marked in blue) and not included in the processed symptoms. Besides, three datasets **Symptoms**, **Complaints**, and **Integrations** are generated to validate the performance of the proposed methods and architectures. The Symptoms dataset only contains the symptoms processed by Step3 (for example processed symptoms in Table 4.2). The Complaints dataset only contains the cleaned complaint text processed by Step1(for example complaint text in Table 4.2). The Integration dataset consists of data generated through Step4 (for example network input in Table 4.2).

4.3.2 The Hybrid architecture

Our hybrid architecture combines a pre-trained encoder-based model (such as BERT, DistilBERT, XLNet or Roberta) and a traditional deep learning model (CNN, LSTM or their variants), which makes full use of the words' features from two different representation spaces. The classification accuracy of CNN and LSTM is similar to the text classification tasks, but CNN is more efficient due to the use of parallel computing. Therefore, this chapter picks CNN from traditional deep learning methods as an example to fuse with the pre-trained encoder-based models. The input of this architecture is the three datasets processed in Section 3.1 and the output is the prediction of the 4 common neurological diseases. This chapter takes BERT+CNN as an example to explain the training process. As depicted in Figure 4.2, the left part is the BERT branch, and the right part is the CNN branch (the Bert branch can be changed to other pre-trained encoder-based models).

Given the neural network input data as $X = (x_1, \ldots x_i, \ldots, x_m)$, four diseases as $Y = (y_1, \ldots, y_4)$ in which x_i is the ith word in the text and m is the length of the text.

BERT Branch

The first hidden state H in Figure 4.2 of the BERT model presents the features of the whole input text, it is used as the input of the lateral linear layer. This chapter utilises the sigmoid function in the softmax layer as

$$Bert_{logits} = Softmax(Linear(H)). \tag{4.1}$$

The Cross-Entropy Loss is used as the loss function, the loss of the BERT branch is described as

$$Loss_{bert} = CrossEntropyLoss(Bert_{logits}, Y). \tag{4.2}$$

CNN Branch

Word embedding is a learned representation for text where words that have the same meaning would have a similar representation. As mentioned in the introduction part, there are mainly four pre-trained word embedding

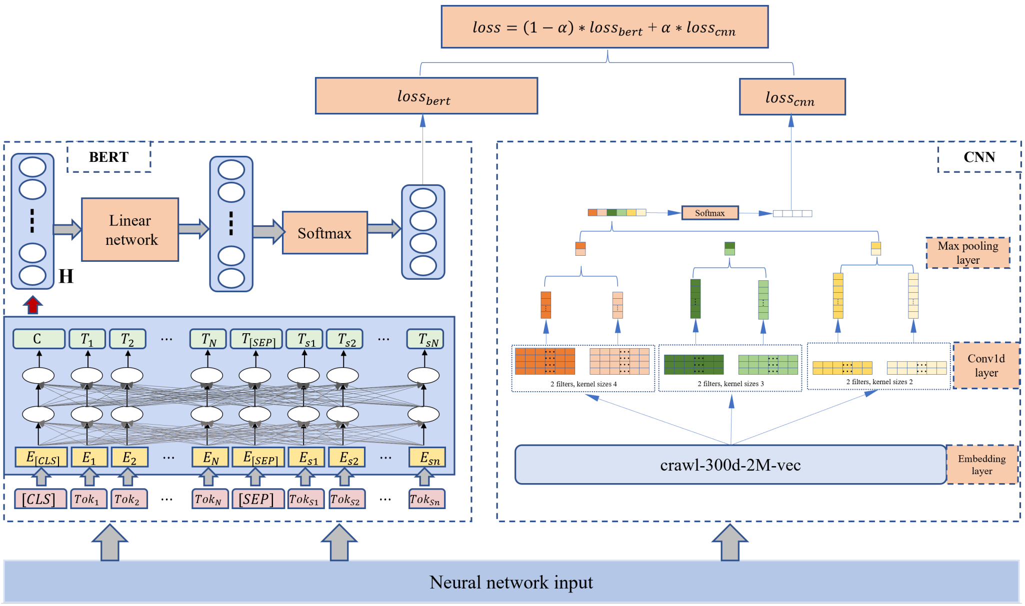


Figure 4.2: The hybrid architecture.

models. Because of the wide coverage of different semantic lexical, the 'crawl-300d-2M.vec' (Mikolov et al. 2017) (2-million-word vectors trained on Common Crawl (600B tokens)) of Google's fastText are utilised as the embedding layer of CNN as shown in Eq (4.3).

$$X_{vector} = Embedding(X). \tag{4.3}$$

The 1-dimension CNN is chosen as the convolutional layer, which is shown below:

$$CNN_{output} = conv1d(X_{vector}). \tag{4.4}$$

The max pooling layer and linear classification layer are shown in Eq (4.5).

$$CNN_{linear} = Linear(Maxpooling(CNN_{output})). \tag{4.5}$$

The softmax layer is described as the following

$$CNN_{logist} = Softmax(CNN_{linear}). \tag{4.6}$$

The loss function of the CNN model is the same as BERT, as shown

$$Loss_{CNN} = CrossEntropyLoss(CNN_{logist}, Y). \tag{4.7}$$

A parameter α for the hybrid architecture is used to fuse the two branches, presented below

$$Loss = \alpha * Loss_{CNN} + (1 - \alpha) * Loss_{Bert}. \tag{4.8}$$

The $(m + 1)th$ fusion weight α_{m+1} is updated by the following equation

$$\alpha_{m+1} = \alpha_m - (Loss_{CNN} - Loss_{Bert}) * lr * \alpha_m. \tag{4.9}$$

In which the parameter lr is the learning rate.

4.3.3 Evaluation Metric

This research uses predicting accuracy as the evaluation metric. Predicting accuracy is a common evaluation metric used to assess the performance of the proposed methods and hybrid architecture. It measures the proportion of correct predictions made by the model out of the total number of predictions. The accuracy is calculated using the following formula:

$$ACC = \frac{TP + TN}{N} \tag{4.10}$$

Where TP (True Positives) is the number of instances that were correctly predicted as positive, TN (True Negatives) is the number of instances that were correctly predicted as negative, and N is the total number of predictions.

4.4 Experiment

4.4.1 Experiment settings

This chapter prepares three training datasets, namely the Symptoms dataset, Complaints dataset and Integration dataset to validate the performance of the proposed architecture and methods. The Symptoms dataset and Complaints dataset consists of processed symptoms and complaints, respectively, and the Integration dataset consists of texts processed by the Complaint-Symptoms Integration Method. To validate the hybrid architecture and the proposed training strategies, the training, validation and testing samples are divided in a ratio of 7:2:1, which has 179/51/27 samples, respectively. Meanwhile, four tasks are arranged using classification accuracy as the metric.

4.4.2 Bert branch baseline

Comparing the performance of different pre-trained encoder-based models on the proposed hybrid architecture, four commonly used sentence classification models are adopted. The details of the four models are as below:

- **BERT** is a pre-trained language model that has achieved state-of-the-art performance in various NLP tasks. It uses a bidirectional transformer architecture, which allows it to consider the entire context of a sentence or text when generating its representations.

- **DistilBERT** is a smaller and faster version of BERT, designed for resource-constrained environments, while still maintaining high accuracy in various NLP tasks. It uses a distillation technique to transfer the knowledge from the larger BERT model to a smaller one.

- **XLNet** is a transformer-based language model that aims to address some of the limitations of previous models like BERT. It uses a permutation-based approach for generating representations, which allows it to capture more diverse and complex relationships between words in a sentence.

- **RoBERTa** (Robustly Optimized BERT approach) is another variant of BERT, which was pre-trained on a larger dataset and with improved training techniques. It has achieved state-of-the-art performance on several NLP tasks, outperforming BERT and other previous models.

Parameters setting: The batch size is set to 8, dropout p=0.5, the optimisation method is Adam (Kingma and Ba 2014) and the learning rate is set as 1e-6. For the CNN branch, this work chooses the 1-dimension CNN which has three kernel sizes (2,3,4) and each kernel has two filters. Besides, the initial value of α for the hybrid architecture is set as 0.3 with lr=1e-5. The hidden unit of CNN uses the fixed embedding representation vector (dimension: 300) and the epoch is set as 200.

The configuration of the computer is Windows 10, Nvidia RTX 2080super, Intel Core (TM) i7-9700 and RAM 32G.

4.4.3 Experimental design

In order to validate the diagnostic accuracy of the proposed methods, three experiments are designed. Besides, three training methods are adopted for each architecture, including training on the symptoms processed with the Symptom Dot Separating Method, training on the complaint text only and training on the data processed with the Complaint-Symptoms Integration Method.

Fine-tuning the pre-trained encoder-based model: This step will further fine-tune the pre-trained encoder-based models such as BERT, DistilBERT, XLNet and Roberta on the neurology referral letter dataset. The architecture of this step can be found in the left BERT branch of Figure 4.2.

The hybrid architecture: This part is designed to compare the accuracy of different combinations between BERT+CNN, DistilBERT+CNN, Roberta+CNN and XLNet+CNN to validate the performance of the proposed hybrid architecture.

Comparison with the sequential model: As shown in Figure 4.3, the sequential model consists of the pre-trained encoder-based network and CNN network, and the pre-trained encoder-based model can be seen as the embedding of CNN. A single embedding representation space is utilised in this architecture, and it is used to compare the effectiveness of the hybrid architecture which fuses two different embedding representation spaces. Four models are designed for this task, including the BERT-CNN, DistilBERT-CNN, Roberta-CNN and XLNet-CNN. The hidden unit of the CNN in this section maintains the identical dimensionality as the pre-trained encoder-based embedding model, which is 768 dimensions.

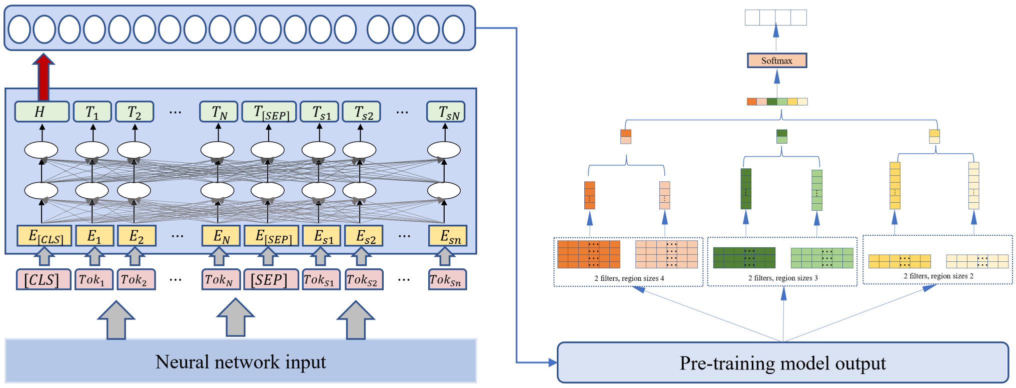


Figure 4.3: Bert-CNN sequential model.

Validation of adding dot: By comparing the performance of adding dots and without dots for the extracted symptoms, this task is designed to validate the effect of the Symptom Dot Separating Method on the proposed hybrid architecture.

4.4.4 Experiment results

Accuracy results of the design tasks on the neurology dataset are presented in Table 4.3. As described in Section 4.4.3, the first vertical column of Table 4.3 shows three designed architectures, and the second column shows different models in each architecture. Symptoms, Complaints and Integration are three training datasets described in Section 4.3.1.

Table 4.3: Accuracy results of the four designed experiments.

Architectures	Models	Symptoms	Complaints	Integration
Fine-tuning pre-trained encoder-based models	BERT	0.795	0.772	**0.846**
	DistilBERT	0.829	0.823	**0.846**
	Roberta	0.782	0.823	**0.833**
	XLNet	0.806	0.759	**0.808**
	CNN	0.821	0.734	**0.846**
Hybrid architectures	BERT+CNN	0.860	0.759	**0.897**
	DistilBERT+CNN	0.846	0.810	**0.872**
	Roberta+CNN	0.822	0.823	**0.872**
	XLNet+CNN	0.822	0.823	**0.885**
Sequential architectures	BERT-CNN	0.760	0.696	0.654
	DistilBERT-CNN	0.829	0.671	0.654
	Roberta-CNN	0.808	0.734	0.795
	XLNet-CNN	0.756	0.772	0.782

As shown in the black bold data of pre-trained encoder-based models in Table 4.3, the classification accuracy of training on the Integration dataset for each model is significantly higher than the results of training on Complaints and extracted Symptoms alone. With an average of 3% and 3.9% improvement on the Symptoms and Complaints datasets respectively. Besides, CNN also has an obvious improvement. It indicates that training on the data processed with the data augmentation methods will produce better classification accuracy. Comparing the result in the Complaints column of pre-trained models with CNN, it can be found that fine-tuning on pre-trained encoder-based models perform better than CNN (as 0.734 is lower than the other results in the complaints column). But the proposed Complaint-Symptoms Integrating training method 're-energizing' CNN, makes it outperform Roberta and XLNet and achieve comparable performance to BERT and DistilBERT.

This also can be evidence of the positive effect of the proposed data integration method.

Two conclusions can be drawn from the hybrid architecture of Table 4.3:

- By comparing the performance of hybrid architectures and their associated pre-trained encoder-based models under the same training dataset (for example comparing BERT+CNN and BERT on the Symptoms, Complaints or Integration dataset), it can be found that on both Symptoms and Integration datasets, all hybrid architectures outperform pre-trained encoder-based models. Besides, BERT+CNN architecture obtains the best classification accuracy on the integration dataset with a 5.1% improvement.

  The training loss and validation accuracy for the BERT+CNN architecture are presented in Figure 4.4. The vertical axis on the left is the label of loss (red line), and the right one is the label of the testing accuracy (blue line). The horizontal axis is the training epochs. It can be seen from Figure 4.4 that the loss and accuracy tend to converge.

- As for the hybrid architectures in Table 4.3, by comparing the classification accuracy of different training datasets under the same architecture, the blue bold data in Table 4.3 shows that all non-sequential architectures benefit from the Integration dataset: models trained on it consistently achieve better performance than either Symptoms or Complaints on their own.

These two conclusions demonstrate the positive effect of the hybrid architecture and two data augmentation methods on neurology disease classification tasks. The proposed method brings a maximum of 11% improvement for disease classification. The reasons for this improved performance are twofold: the two data augmentation methods enrich the content of training data, and they combined the original complaint text with the extracted symptoms. This highlights the key information in the complaint text. Another is that the hybrid architecture fuses two embedding spaces, which enriches the representation of the collected neurology dataset. Therefore, training on them results in better classification accuracy.

Comparing the classification accuracy of the sequential architectures and their related hybrid architectures (such as BERT+CNN and BERT-CNN) in Table 4.3, the proposed pre-trained encoder-based models+CNN outperforms pre-trained encoder-based models-CNN on all three training datasets. This is because the proposed architecture fuses two different embedding representation spaces, which can improve the accuracy of neurological disease classification tasks significantly.

Table 4.4: Validating the performance of the Symptom Dot Separating Method.

Conditions	Symptoms	Symptoms + Complaints
Without dot	0.829	0.846
With dot	**0.860**	**0.897**

Due to the superior performance of the hybrid architecture BERT+CNN, it is chosen to validate the performance of the Symptom Dot Separating Method, the result can be found in Table 4.4. This table shows that adding dots between each symptom improves the classification accuracy of the proposed hybrid architecture. The main reason is that each symptom is separated, without semantic meaning, adding dots can avoid semantic confusion.

Summary This chapter proposed three methods, including the hybrid architecture, two data augmentation methods Symptom Dot Separating Method and the Complaint-Symptoms Integration Method which are proven to have positive effects on the accuracy of neurological disease classification tasks. This means that it can be a good assistant tool for GPs to make the primary diagnoses.

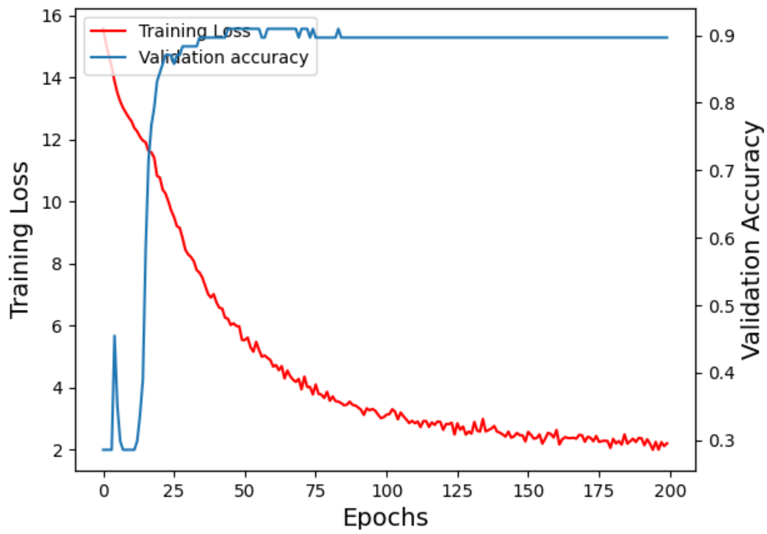


Figure 4.4: The training loss and validation accuracy for the BERT+CNN architecture.

4.5 Application

In order to help GPs complete diagnosis more efficiently and accurately, this work proposed the hybrid architecture to develop an AI diagnosis assistant web application, which can help GPs make a primary diagnosis by talking/-texting with them. Flowchart (left part in the figure) and demo screenshot of the conversation between a GP and AI diagnosis assistant (right part in the figure) can be found in Figure 4.5. The natural language understanding part in Figure 4.5 follows the same steps as the data processing in section 4.3.1. Besides, both Speech-to-text (STT) and Text-to-speech (TTS) functions are implemented using state-of-the-art methods, which can better improve the experience of this AI diagnosis assistant web application. A demonstration video which shows the process of both writing and speaking assistant process can be found at `https://www.youtube.com/watch?v=Coj1xGYOCBw`.

4.6 Summary

This chapter introduces a hybrid architecture, targeting to help GP to improve the accuracy of primary diagnosis. Comparative experiments on the

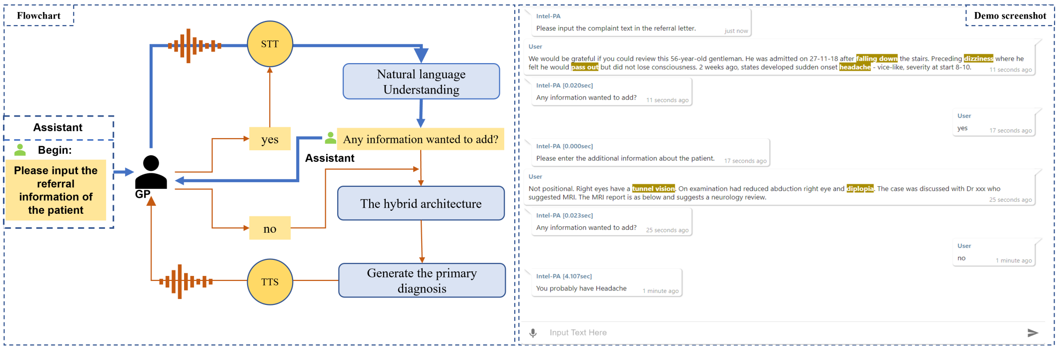


Figure 4.5: The flowchart and demo of the assistance diagnosis application which is developed on the hybrid architecture.

neurology dataset demonstrate that the proposed model has satisfactory classification accuracy. Two data augmentation methods (the Symptom Dot Separating Method and the Complaint-Symptoms Integration Method) which were proved effective in improving the predicting accuracy of the neurology disease are also be proposed. This paper also applied the hybrid architecture to develop an AI diagnosis assistant web application, which can efficiently help GPs make a primary diagnosis by talking/texting with them. The accuracy of this architecture needs to be improved, and in the future, the relation between each symptom will be presented via the knowledge graph. The dataset and the source code can be found at the GitHub link which is shared in the Introduction section. It is worth mentioning that owing to the development schedule and the intricate policy of the collaborated hospital, the evaluation of the developed web applications in this chapter—in genuine clinical contexts is presently under development by my colleagues.

Chapter 5

Capability of Large language model on assisting GPs for diagnosis

5.1 Background

With the evolution of global healthcare, the hierarchical healthcare system has emerged as an advanced model, adopted by numerous countries (Jia et al. 2017, Tao and Han 2021, Gillam 2002, Goldfield et al. 2003, Liang et al. 2022). Within this hierarchical framework, GPs assume a pivotal role as the initial point of patient contact, playing a crucial part in primary care procedures (O'Donnell 2000). GPs are tasked with addressing a wide array of general patient issues during the first consultation, ensuring prompt and preliminary diagnoses. In instances where primary care is unable to address the specific needs of patients, or when an expert opinion is deemed necessary, patients are directed to specialised medical practitioners (Ramanayake 2013).

In the context of the COVID-19 pandemic, the NHS of England persistently grapples with significant operational strains. The capacity of secondary care confronts substantial challenges in accommodating patients referred by GPs, resulting in extended waiting periods and, in some instances, rejections of referrals. Such situations compromise patient welfare and overall health. Data pertaining to "no slots" associated with Appointment Slot Issues (ASIs) in the NHS e-Referral system—which processes approximately

95% of GP referrals in England—sheds light on the capacity constraints experienced when attempting to schedule referral appointments. This chapter has analysed the "no slots" data from ASIs reports spanning January 2017 to July 2023, as illustrated in Figure 5.1. Figure 5.1 indicates that the "no slots" situation within the NHS remained relatively stable, registering below 280,000/month prior to June 2020. However, post-June 2020, the "no slots" number surged past 350,000/month, maintaining this elevated level up to July 2023. The instances of referrals without available slots escalated from 270,548 in June 2020 to 401,692 by July 2023, marking a 48% increase. When juxtaposed with data prior to the COVID-19 pandemic, such as the 168,407 recorded in July 2017, there is a staggering increase of 138.5%. This underscores a pressing challenge that the NHS must address.

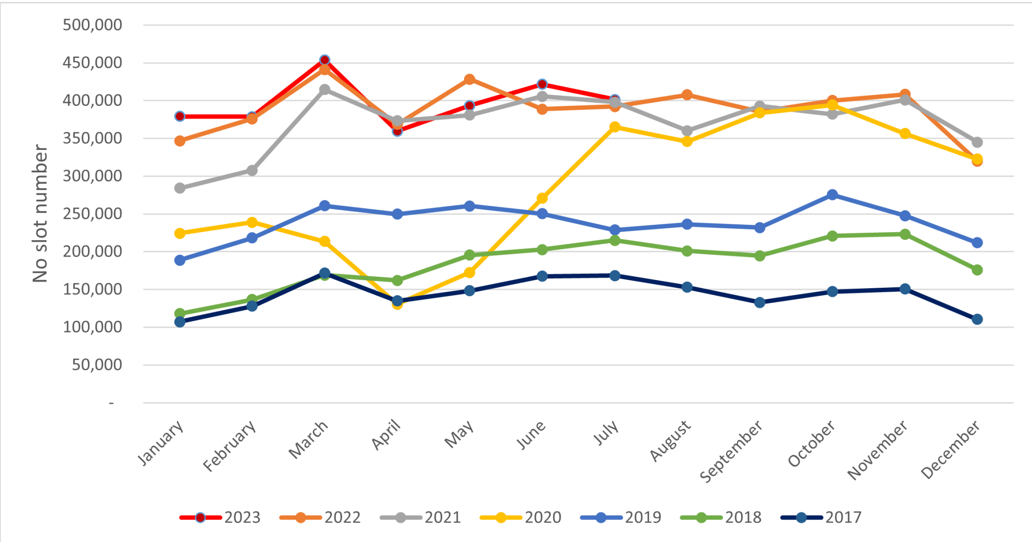


Figure 5.1: The distribution of "no slots" for GP referrals within NHS England from January 2017 to July 2023.

As the first point of contact for patients, GP has the best understanding of the patient's situation. If there exists an assistant can help GPs to make professional and precise diagnoses, along with expert guidance, the rate of referrals could potentially decrease. This, in turn, would alleviate the burden on specialists and ensure that patients really requiring secondary care receive expedited services. Extensive research efforts have been dedicated to the advancement of AI-assisted diagnostic methods (Lewenberg et al. 2017, Shim et al. 2018, Wei et al. 2018) and tools, such as Symptomate, Patient, HealthTap, Ada, etc. However, these resources primarily target the patient perspective not for GPs. Given that GPs possess a comprehensive understanding of their patients' situations, they are adept at efficiently extracting essential information during patient interactions. They skillfully distil pertinent details, encompassing medical history and current symptoms, into referral letters whenever they seek expert guidance from secondary care providers. Recognising that these referral letters contain rich medical information of patients,

the concept emerges to employ them for training AI-assisted diagnostic tools. This approach aims to assist GPs in making accurate diagnostic decisions by leveraging the content of referral letters.

Although referral letters may contain rich patient medical information, their free-text structure and strict privacy policies necessitate the involvement of professional doctors for processing data, such as anonymization and re-evaluating the labels. The processing of such a referral letters-based dataset is a substantial workload, both time-consuming and costly. Consequently, due to this burdensome task, most doctors lack the extra time required for data labelling, resulting in limited dataset availability. In the existing literature, there is no research on utilising referral letters to assist clinicians in their decision-making processes. To the best of our knowledge, this study is the first to embark on this endeavour. Since the referral letter contains rich free-text information, the NLP techniques can be used for the downstream tasks. However, the effectiveness of NLP heavily relies on the quality and quantity of the training data. The scarcity of training data is a prevalent challenge. This scarcity can pose difficulties in training a precise NLP model capable of generalising effectively to new, unseen samples. Addressing the shortage of training data for disease classification stands as a pivotal concern. Simultaneously, devising novel, efficient, and accurate training architectures for the available data emerges as a crucial pathway in advancing disease prediction based on referral letters.

In recent years, significant advancements have been witnessed in large language models (LLMs), showcasing an extraordinary proficiency in comprehending and generating sequences across an extensive spectrum of domains. These encompass question answering, essay composition, code generation, and computer vision (Brown et al. 2020). LLMs, denoting language models comprising hundreds of billions (or more) of parameters, undergo training on vast volumes of textual data (Shanahan 2022). LLMs exhibit the capacity to extend their skills to uncharted tasks for which they haven't been explicitly trained. This phenomenon, known as few-shot learning or zero-shot learning, highlights the remarkable adaptability of LLMs. This adaptability can be attributed in part to multitask learning, a mechanism that inadvertently allows LLMs to accumulate knowledge from implicit tasks embedded within their training corpus (Radford et al. 2019). The successful implementation of ChatGPT and GPT-4 (OpenAI 2023) in real customer scenarios propels the zero/one/few-shot learning of LLMs into an entirely new phase.

This also brings encouraging developments for the AI+Medical field. With limited available data, the remarkable achievements of ChatGPT and GPT-4 in zero/one/few-shot learning tasks have motivated researchers to explore applications related to medical. The evaluation conducted on the USMLE (Nori et al. 2023) and MultiMedQA (Singhal et al. 2022) datasets reveals that the in-context learning capabilities of ChatGPT and GPT-4 yield significant performance on these two datasets. However, when assessing the efficacy of ChatGPT and GPT-4 in Named-Entity-Extraction (NER) tasks using the NCBI (Doğan et al. 2014) and BC5CDR (Li et al. 2016a) datasets and in Relation Extraction (RE) tasks with the GAD (Rouillard et al. 2016) and EU-ADR (Van Mulligen et al. 2012) datasets, their in-context learning performance lags behind current state-of-the-art methods. Furthermore, the feasibility of employing these models in real-world clinical settings for GPs has yet to be investigated. Therefore, in collaboration with the neurologists from the University Hospital Dorset (UHD), this study collected a labelled dataset comprising referral letters. This dataset serves as the foundation for a comprehensive assessment. This chapter conducted three distinct tasks to validate the efficacy of LLMs in supporting GPs' diagnostic decision-making. The contributions of this chapter are outlined as follows:

Firstly, given that ChatGPT and GPT-4 (OpenAI 2023) exhibit state-of-the-art performance across a wide range of general downstream tasks and both models provide an open API, the current endeavour involves the utilisation of in-context learning methods to evaluate their performance in the domain of referral letter based disease diagnosis. The outcomes of the experiments indicate that the disease prediction accuracy achieved by ChatGPT and GPT-4, through the application of prompt engineering, falls short of being sufficient for aiding doctors in disease diagnosis.

Secondly, Existing research (Feng et al. 2021, Bayer et al. 2022) proved that leveraging LLMs can significantly enhance a dataset by generating new samples that exhibit similar semantic meanings. Therefore, to solve the limitation of low-scale datasets for deep learning methods, this study have employed the ChatGPT model to enhance the comprehensiveness of the referral letter dataset. In the scope of this study, the objective is to employ ChatGPT, a prominent language model, for the purpose of data augmentation. ChatGPT, built upon the GPT-3 architecture (Brown et al. 2020), has undergone training on extensive web data, thereby building upon a diverse and comprehensive knowledge base. Moreover, the training of ChatGPT is distinguished by its utilisation of Reinforcement Learning from Human Feedback (RLHF). This approach involves integrating human feedback into the process of result generation and selection. Specifically, a reward model is established based on rankings provided by human annotators or on generated outcomes. Subsequently, this reward model reinforces model outputs that closely align with human preferences and values. These innovative techniques collectively position ChatGPT as an optimal candidate for producing high-quality human-like data samples. Building upon the groundwork laid by Alpaca (Taori et al. 2023) and ChatAug (Dai et al. 2023), this study utilises the text generation capabilities of ChatGPT to amplify the scope of the referral letters using prompt engineering. Through this methodology, the distinctiveness of the generated dataset within the embedding space in comparison to actual referral letters is examined. The analysis extends to investigating the potential duplication within the data distribution, particularly addressing concerns related to overfitting. And this study made the augmented dataset public for future researchers.

Finally, the generated datasets are harnessed to facilitate the training of models intended for subsequent disease classification endeavours, thereby substantiating the efficacy of the generated dataset. This phase employed fine-tuning on two distinct types of models: firstly, due to transformer (Vaswani et al. 2017) encoder-based models are mostly used for text classification tasks; therefore, the generated dataset on the encoder-based Pre-trained Language Models (PLMs) such as BERT (Devlin et al. 2018a), XLNet (Yang et al. 2019b), RoBERTa (Liu et al. 2019a), and BioBert (Lee et al. 2020) are tested for disease classification.

The experiment shows that the prompt engineering accuracy of the ChatGPT and GPT-4 is not good enough for GPs to use in a real clinical scene. Since LLMs are generated models and based on the decoder of transformer architecture, they are seldom used for text classification tasks. But the performance of the LLMs still tempts us to try more on the necessary tasks. Besides, considering the safety and privacy concerns caused by directly uploading patient information onto the LLM's platform through their APIs in the real clinical scene (Aziz 2017), this study fine-tuned the open-source LLMs on the dataset locally. Constrained by hardware and cost, the testing was confined to the 7B and 13B Llama models (Touvron et al. 2023). In this scene, the referral letter classification is treated as a multiple-choice question-answer (QA) task, with the target diseases being treated as options within the provided referral letter text. The model's output becomes the predicted option. Notably, the findings demonstrate that, when considering fine-tuning with the same data, the pre-trained encoder-based model outperforms the LLMs on the referral letter-based disease classification task.

In conclusion, considering the NHS's existing burdens, as the first one who use referral letters as a training dataset, the aim of this chapter is to ascertain an optimal strategy to assist GPs in decision-making by leveraging referral letters and to assess the capability of LLMs in referral letter-based disease classification. The contributions of the work are as follows:

• Investigate the performance of zero/one/few-shot approaches using ChatGPT and GPT-4 for disease prediction. Experiments show that the performance of in-context learning for these two models are not satisfactory for real clinical scenarios.

• To overcome the limitations of the small-scale training dataset, this study leveraged ChatGPT's text generation capabilities to augment the dataset. The experiment shows that this augmentation can effectively improve the performance of downstream tasks.

• Utilising the augmented referral letter datasets, this study designed two downstream disease classification fine-tuning solutions, supervised classification on the encoder-based PLMs and multiple-choice based question-answering on LLMs. Experiments demonstrated that encoder-based models significantly outperformed decoder-based LLMs in referral letter-based disease classification, suggesting the potential superiority of encoder-based models for disease diagnosis classification tasks.

5.2 Related work

This section reviews the literature related to the topic of the work, including previous research on medical decision-making, data augmentation and LLMs.

5.2.1 Medical decision-making

Since Ledley and Lusted (1959) published their work on medical diagnosis, numerous computational models and reasoning techniques have been developed to aid clinicians. These encompass tree-based models (Kononenko 1993 2001, Xu et al. 2013, Kohavi et al. 1996, Nan et al. 2015, Hayashi 1990), foundational probabilistic and decision-theoretic frameworks (Gorry and Barnett 1968, Heckerman et al. 1992), rule-based production systems (Shortliffe 1977, Buchanan and Shortliffe 1984), and semantic networks (Patil et al. 1981). Notably, many techniques from this era were rooted in statistical theory, with machine learning and deep learning methodologies yet to be incorporated. Consequently, their performance was intrinsically constrained.

Although subsequent work has harnessed supervised learning techniques using medical databases (Wiens et al. 2016, Henry et al. 2015, Escobar et al. 2020, Caruana et al. 2015) and deep neural network architectures (Lewenberg et al. 2017, Shim et al. 2018, Wei et al. 2018, Xu et al. 2019, Suresh et al. 2017), these have largely been tailored to specific clinical tasks and do not necessarily apply to the diagnosis based on referral letters. Concurrently, the emergence of PLMs and LLMs has shown promise in specialised medical domains, as demonstrated by platforms such as CancerGPT (Li et al. 2023), Med-PaLM (Singhal et al. 2022), and SynerGPT (Edwards et al. 2023). However, the efficacy of LLMs in classifying diseases based on actual clinical referral letters remains a nascent area of exploration. Motivated by the potential of PLMs and LLMs in medical tasks, this chapter proposed different methods to assess their performance in assisting GPs with decision-making.

5.2.2 Data augmentation

Data augmentation is a method employed to enlarge the pool of training data accessible to machine learning models without necessitating additional human data annotation. Expanding the training dataset, assuming it encompasses a reasonable degree of diversity, is crucial for fostering model generalization, particularly in scenes with limited resources.

The Easy Data Augmentation (EDA) (Wei and Zou 2019) approach is predicated on word substitution. Back translation (Sennrich et al. 2015) involves translating text into another language and subsequently reverting it to the original, thereby enriching the linguistic representation's diversity. Encoder-based PLMs such as BERT (Devlin et al. 2018a), BART (Lewis et al. 2019), and CBERT (Wu et al. 2019) have been employed for text data augmentation. Traditional methods predominantly leverage machine learning models to intensify the text's diverse representation. However, given that the representation and semantics of the text can be constrained, the efficacy of these augmentation strategies may be bounded.

Exploiting the text generation capabilities inherent in decoder-based PLMs, both GPT-2 (Radford et al. 2019) and GPT-3 (Brown et al. 2020) emerge as robust tools for data augmentation. Nevertheless, due to the magnitude and data quality constraints of these models, their output remains unpredictable. With an expanded training dataset and the remarkable performance exhibited by ChatGPT and GPT-4 in zero/one/few-shot learning, platforms such as ChatAug (Dai et al. 2023), Aplaca (Taori et al. 2023), and ZeroShotDataAug (Ubani et al. 2023) harness the prompted text generation feature of ChatGPT. Their implementation in generated downstream tasks has demonstrated commendable performance.

5.2.3 Large Language Model

In contemporary research, LLMs are often characterized by their extensive parameter sets, typically in the order of hundreds of billions or more. These models are trained on vast text corpora (Shanahan 2022). Notable examples include GPT-3 with approximately 175B parameters (Brown et al. 2020), PaLM with around 540B parameters (Chowdhery et al. 2022), Galactica with close to 120B parameters (Taylor et al. 2022), and LLaMA, which ranges from 7B to 65B parameters (Touvron et al. 2023). These LLMs are fundamentally based on the Transformer architecture (Vaswani et al. 2017), incorporating deep neural networks with stacked multi-head attention layers. While there is a structural resemblance between LLMs and their smaller counterparts, both in terms of architecture (primarily the Transformer) and pre-training objectives (essentially language modeling), the distinguishing factor for LLMs is their significant scale in model size, pre-training data, and computational requirements. This upscale allows LLMs to demonstrate enhanced proficiency in understanding natural language and producing contextually relevant, high-quality text responses to given prompts. This amplification in capacity is, to an extent, encapsulated by the scaling law, which posits that model performance generally scales with an increase in model size (Kaplan et al. 2020). To elucidate the evolution and distinct features of recently proposed large language models, this study has systematically compiled and analysed statistical data pertaining to these models. This analysis aims to provide a comprehensive overview, emphasizing both the developmental trajectory and individual

Table 5.1: Details of released LLMs Subsequent to GPT-3

LLMs Name	Release date	Institute	Language	Param	Open source
GPT-3 (Brown et al. 2020)	2020-05	OpenAI	English	175B	Yes
LaMDA (Cheng and Thoppilan 2022)	2021-05	Google	English	137B	No
Jurassic-1 (Lieber et al. 2021)	2021-08	AI21	English	178B	No
MT-NLG (Smith et al. 2022)	2021-10	Microsoft, NVIDIA	English	530B	No
Gopher (Rae et al. 2021)	2021-12	DeepMind	English	280B	No
InstructGPT (Ouyang et al. 2022)	2022-03	OpenAI	English	1.3-175B	No
Chinchilla (Hoffmann et al. 2022)	2022-04	DeepMind	English	70B	No
PaLM (Chung et al. 2022)	2022-04	Google	Multi-Lingual	540B	No
OPT (Zhang et al. 2022a)	2022-05	Meta	English	125M-175B	Yes
BLOOM (Scao et al. 2022)	2022-11	BigScience	Multi-Lingual	176B	Yes
Galactica (Taylor et al. 2022)	2022-11	Meta	English	120B	Yes
GLM-130B (Zeng et al. 2022)	2022-10	Tsinghua	CN, EN	130B	Yes
LLaMA (Touvron et al. 2023)	2023-02	Meta	Multi-Lingual	7B-65B	Yes
GPT-4 (OpenAI 2023)	2023-03	OpenAI	Multi-Lingual	~ 100 T	No

characteristics of each model, thus fostering an accessible understanding of their respective capabilities and limitations within the broader scientific and technological context. The summarization of the recently released LLMs can be found in Table 5.1.

5.3 Dataset

In collaboration with the neurologist from the NHS Trust, this study obtained a collection of GP referral letters for the patients. An exemplar of this referral letter is presented in Table 5.2. Within the disease diagnosis column, the terms R56.9, M54.5, G40.3, and G43.9 represent ICD-10 codes, with the subsequent text denoting the conventional nomenclature for the diseases associated with each respective code. It should be noted that, in the original referral letters, the disease names corresponding to the ICD-10 codes might differ due to the individual professional and personal preferences of the GPs. To ensure consistency and accuracy in nomenclature, this study referenced the official database at `https://www.icd10data.com/` to align the disease names with their respective ICD-10 codes.

Table 5.2: The examples of referral letters.

Complaint text	Disease diagnosis
Dear Dr XXXX, We would be grafeful if you could review Mr XXXX in one of your Rapid Acsess Neurology Clinic. He was seen on Ansty under Dr XXXX with a history of severe sudden onset headache followed by visual disturbances and unsteadyness within minutes. His visual disturbance is unusual, describing bilateral diplopia which persists as bi and monocular vision. Initially he also described very restriced tunnel vision that have steadily improved during his admission but still not returned to normal. On examination he has normal ocular movements, symmetrical tunnelling of the visual fields in both eyes. His pupillary reflexes and fundoscopy are normal. No other abnormal neurology was detected.	R56.9 - Unspecified convulsions
39 year old lady was involved in a road traffic accident about two weeks ago when her car was shunted from behind and she experienced some low back pain. She felt it was getting gradually better but in the last 48 hours the pain has got worse and has been associated with urinary incontinence and faecal incontinence. Getting some tingling in both legs.She had an MRI scan of her spine last night but it was completely clear. Her problem does not seem to be due to any spinal problem. Discussion with general surgeons: informed that her presentation would not constitute a general surgical emergency and that the patient should be referred to a pelvic floor surgeon via GP. Medical registrar advised that all back pain patoients are to reviewed by rheumatology team and no medical team imput is necessary. Discussed with rheumatology consutant: no need for rheumatology review, reffer to Neurology tweam.	M54.5 - Low back pain

17 year old with first episode of seizure PC - seizure at work XX-XX-XXXX morning - customer noted eyes roll back, became rigid, fell backwards, shaking arms and legs for 2 minutes. Tongue biting, no incontinence. 1 minute later, had another similar episode. Postictal phase - can't remember going to work. Later that night, attended party - had excess alcohol, friend stated knocked back whilst sitting on curb and hit her head on pavement, no immediate loss of consciousness.	G40.3 - Generalised idiopathic epilepsy and epileptic syndromes
We would be grateful if you could see this 35 year old lady as a inpatient on XXX. She has previously been seen by Dr XXXX with BIH and had a VP shunt inserted on the XX-XX-XXXX. She required adjustment to this on XX-X-XXXX as VP shunt was in subcutaneous fat instead on peritoneum. She has presented with a week history of worsening headache and neck pain with associated blurred vision. On examination she has bilateral papilloedema. CT and shunt series have been performed and reviewed by SOTON. Initially they suggested proceeding to LP which unfortunately was unsuccessful and Anaesthetic assistance has been requested on CEPOD. SOTON have been back in contact and suggest ophthalmology review first. We would be grateful for your assessment of this lady with her management and treatment.	G43.9 - Migraine, unspecified

In the original referral letters, a total of 17 diseases were identified, encompassing 439 cases. The distribution of these cases across the diseases is detailed in Table 5.3, with a visual representation provided in Figure 5.2.

Table 5.3: Case number for the 17 diseases.

Disease name	Number
A41.9 - Sepsis, unspecified organism	16
G35 - Multiple sclerosis	34
G40.3 - Generalised idiopathic epilepsy and epileptic syndromes	18
G40.9 - Epilepsy, unspecified	20
G43.9 - Migraine, unspecified	35
G61.0 - Guillain-Barre syndrome	17
I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified	12
I63.9 - Cerebral infarction, unspecified	30
J18.9 - Pneumonia, unspecified organism	13
J69.0 - Pneumonitis due to inhalation of food and vomit	12
M54.5 - Low back pain	9
N39.0 - Urinary tract infection, site not specified	15
R29.8 - Other symptoms and signs involving the nervous and musculoskeletal systems	24
R41.0 - Disorientation, unspecified	11
R51 - Headache	50
R55 - Syncope and collapse	36
R56.9 - Unspecified convulsions	87

5.4 Methods

This section delineates three fundamental components of the study. Initially, this study explores the in-context learning of ChatGPT and GPT-4 for disease diagnosis. Within this framework, two distinct prompt mechanisms are introduced : the first entails directly requesting the diagnostic result based on the information provided in the referral letter (Section 5.4.1.1), and the second involves direct diagnosis followed by the selection of the correct answer from a set of given options (Section 5.4.1.2). Subsequently, elaborating on the

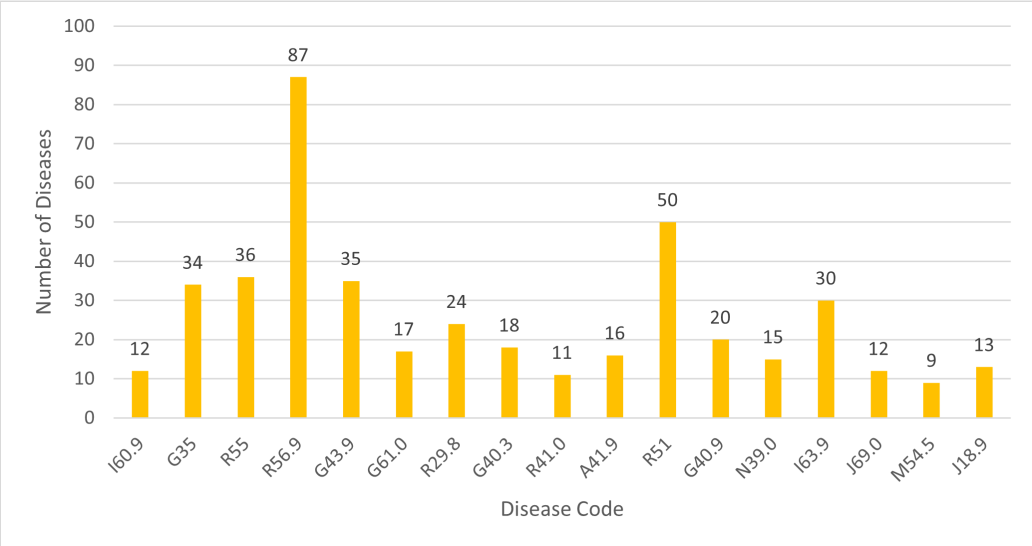


Figure 5.2: Distribution of disease cases

utilisation of ChatGPT to enhance the collected referral letter dataset (Section 5.4.2). This process involves employing prompt engineering techniques, and presenting the evaluation metrics used to assess the quality of the augmented dataset (Section 5.4.3). Finally, applying the enhanced dataset to a downstream task for disease diagnosis (Section 5.4.4). In this phase, this study detail the fine-tuning of encoder-based PLMs and decoder-based LLMs on the augmented dataset, respectively.

5.4.1 In-context learning for disease diagnosis using ChatGPT and GPT-4

The performance of ChatGPT and GPT-4, particularly in in-context learning and logical reasoning, is notably outstanding. This study aims to evaluate their proficiency in disease diagnosis based on referral letters to assist GPs. In partnership with neurologists from UHD, this study amassed a labelled referral letter dataset, which served as the ground truth for the following evaluation. This research employs two methodologies: direct diagnosis and a multiple-choice question-and-answer approach, both of which are elaborated upon subsequently.

5.4.1.1 Direct diagnosis

This section aims to assess the efficacy of the GPT model in direct diagnosis through three distinct prompt engineering methodologies: the zero-shot process and the one-shot process. The "direct diagnosis" refers to a procedure in which the model is tasked with predicting the specific disease name or its corresponding ICD-10 code based on the given information. Prior to delving into the details of this process, it is imperative to define the associated variables. Let $r_i \in R(i = 1 \dots n)$ represent the complaints in a referral letter, where R is the set of referral letters from the collected dataset, and n denotes the total number of these letters. Similarly, the diagnosed disease name is represented as $d_j \in D(j = 1 \dots m)$, where D constitutes the set of disease names and m is the count of distinct diseases. This section introduces a specific Instruct, phrased as: "Please provide only the ICD-10 code and the official name of the diagnosed disease as stated in the corresponding referral letter." The predicted disease name or ICD-10 code, associated with a given referral letter r_i and its true diagnosis d_j, can be designated as p_i. Lastly, the model utilised for this research is symbolised as GPT, encompassing both ChatGPT and GPT-4. The input instances employed for the GPT model through the utilisation of three prompt engineering methods are elucidated in Table 5.4.

In the context of a zero-shot process, the GPT model does not need to train with annotated data, allowing for direct utilisation in predicting disease labels for a given referral letter. The zero-shot process is defined as:

$$p_i = GPT(Instruct, r_i) \tag{5.1}$$

A one-shot process is when a model learns to recognise a new category or task by using only one example. The one-shot process is defined as:

$$p_i = GPT(Instruct, example, r_i) \tag{5.2}$$

where the example is a pair consisting of an authentic referral letter and its corresponding diagnostic outcome, which can be found at Table 5.4 and is defined as:

$$example = [r_k, d_j] \tag{5.3}$$

The prediction accuracy is used as a metric, which is defined as:

$$acc = \frac{sum(f(p_i, d_j))}{n} \tag{5.4}$$

$$f(p_i, d_j) = \begin{cases} 1 & \text{if } p_i = d_j \\ 0 & else \end{cases} \tag{5.5}$$

Furthermore, based on prior experience with referral letter classification using decoder-based PLMs, the symptoms articulated in the referral letters have been observed to significantly influence prediction accuracy. To evaluate the efficacy of associating extracted symptoms with their corresponding referral letters in enhancing prediction accuracy with ChatGPT and GPT-4, the extracted symptoms with the relevant referral letter text are integrated using the delimiter "[SEP]". The input example of this process can be found at the "one-shot-symp" row in Table 5.4 and extracted symptoms from the referral letter text are highlighted in red.

This section employed the symptom-associated referral letter, denoted as r_{symp}, to substitute r_i in the above equation. r_{symp} can be expressed as:

$$r_{symp} = r_i[SEP]s_1[SEP]s_2 \ldots [SEP]s_k \tag{5.6}$$

Table 5.4: Prompt examples for the direct diagnosis process.

Prompt type	Prompt examples
Zero-shot	Please provide only the ICD-10 code and the official name of the diagnosed disease as stated in the following referral letter. Referral letter: Presented with tonic clonic seizure, self-resolved lasted approximately 2-3 minutes Reports had no warning of episode, doesn't remember period after. Remembers coming in the ambulance. Now memory poor, increased confusion AMTS: 5-10. Not at baseline-daughter reports normally happens after seizure. Has had increased episodes of passing out Had similar episodes before, on sodium valproate Not been seen by neuro before BG: dementia ETOH excess, no withdrawal symptoms. Diagnosis:

One-shot	Please provide only the ICD-10 code and the official name of the diagnosed disease as stated in the following referral letter. Referral letter: Presented with tonic clonic seizure, self-resolved lasted approximately 2-3 minutes Reports had no warning of episode, doesn't remember period after. Remembers coming in the ambulance. Now memory poor, increased confusion AMTS: 5-10. Not at baseline-daughter reports normally happens after seizure. Has had increased episodes of passing out Had similar episodes before, on sodium valproate Not been seen by neuro before BG: dementia ETOH excess, no withdrawal symptoms. Diagnosis: G409 - Epilepsy, unspecified Referral letter: Chest pain and collapse x2 (once in ED, hit head on 1st seizure) Impression - Generalised seizures secondary to old stroke. Please can you come and assess the patient and advise on which Antiepileptic medication can be started . Diagnosis:

One-shot-symp	Please provide only the ICD-10 code and the official name of the diagnosed disease as stated in the following referral letter. Referral letter: Presented with tonic clonic seizure, self-resolved lasted approximately 2-3 minutes Reports had no warning of episode, doesn't remember period after. Remembers coming in the ambulance. Now memory poor, increased confusion AMTS: 5-10. Not at baseline-daughter reports normally happens after seizure. Has had increased episodes of passing out Had similar episodes before, on sodium valproate Not been seen by neuro before BG: dementia ETOH excess, no withdrawal symptoms. [SEP]tonic clonic seizure [SEP] memory poor [SEP] confusion [SEP] seizure [SEP] passing out [SEP] dementia. Diagnosis: G409 - Epilepsy, unspecified. Referral letter: Chest pain and collapse x2 (once in ED, hit head on 1st seizure) Impression - Generalised seizures secondary to old stroke. Please can you come and assess the patient and advise on which Antiepileptic medication can be started [SEP]Chest pain [SEP] collapse [SEP] hit head [SEP] seizure [SEP] Generalised seizures [SEP] stroke. Diagnosis:

5.4.1.2 Multiple-choice question and answering

Contrary to direct diagnosis methods, the proposed multiple-choice question and answer technique incorporates all disease names as options within the model's prompt text. This method invariably produces a specified choice as its output. As the output is predetermined, the accuracy of disease predictions may surpass that of direct diagnoses. Analogous to the previously described zero/one-shot methodologies, the diagnostic process can be formulated as:

The zero-shot process:

$$opt_i = GPT(Definition, r_i, [question, option_{set}]) \tag{5.7}$$

The one/few-shot process:

$$opt_i = GPT(Definition, [r_k, question, option_{set}, d_j]_{few}, [r_i, question, option_{set}]) \tag{5.8}$$

Where:

- Definition is phased as "For this task, we request that you make a diagnosis by examining the provided referral letter."

- $option_{set}$ represents the set of disease name options.

- opt_i denotes the predicted choice within $option_{set}$.

- The variables GPT, r_i, and d_j are consistent with the parameters defined in the direct diagnosis method.

- $question$ serves as the directive for the model's output.

For illustrative purposes, an example of a multiple-choice question and answer prompt is provided in Table 5.5. It is noteworthy that, for the few-shot process, five samples were chosen to help the GPT model understand the paradigm of the prediction task. And to make Table 5.5 more clear, the samples associated with the one-shot and few-shot processes have been designated with the colour blue. This deliberate distinction serves to differentiate these samples from the reference letter employed for prediction purposes.

Table 5.5: Prompt example of a multiple-choice question and answering method.

Prompt type	Prompt examples
Zero-shot	Definition: For this task, we request that you make a diagnosis by examining the provided Referral letter. Referral letter: Dear Dr XXXX, We would be grafeful if you could review Mr XXXX in one of your Rapid Acsess Neurology Clinic. He was seen on XXX under Dr XXXX with a history of severe sudden onset headache followed by visual disturbances and unsteadyness within minutes. His visual disturbance is unusual, describing bilateral diplopia which persists as bi and monocular vision. Initially he also described very restriced tunnel vision that have steadily improved during his admission but still not returned to normal. On examination he has normal ocular movements, symmetrical tunnelling of the visual fields in both eyes. His pupillary reflexes and fundoscopy are normal. No other abnormal neurology was detected. Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalised idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine, unspecified (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 - Pneumonia, unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 - Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collaps (Q) R56.9 - Unspecified convulsions. Output:

One-shot	Definition: For this task, we request that you make a diagnosis by examining the provided Referral letter. Referral letter: This gentleman was admitted with SUO after a recent discharge from RBH for rehab 2-52. She has a background of RA and polyarthralgia, which has been worsening over 6-12. Had transient R sided weakness XX-XXXX. We would appreciate a neurology opinion on her globally increased tone and decreased mobility. Thank you. Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalised idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine, unspecified (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Non-traumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 - Pneumonia, unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 - Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collaps (Q) R56.9 - Unspecified convulsions. Output: A Referral letter: Mr XXXX, 72 year-old gentleman, was admitted post-radical chemo-radiotherapy for SCC oesophagus with reduced mobility and poor oral intake. On assessment we are concerned that some of these symptoms may be related to possible Parkinsonism. Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalised idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine, unspecified

(F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 - Pneumonia, unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 - Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collaps (Q) R56.9 - Unspecified convulsions.

Output:

Few-shot	Definition: For this task, we request that you make a diagnosis by examining the provided Referral letter. Referral letter: This gentleman was admitted with SUO after a recent discharge from RBH for rehab 2-52. She has a background of RA and polyarthralgia, which has been worsening over 6-12. Had transient R sided weakness XX-XXXX. We would appreciate a neurology opinion on her globally increased tone and decreased mobility. Thank you. Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalized idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 – Pneumonia unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 – Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collapse (Q) R56.9 - Unspecified convulsions. Output: A Referral letter: Repat from SGH: Previous subarachnoid hemorrhage XX-XX-XXXX. Aneurysm was coiled in SGH on the XX-XX-XXXX. She developed a focal seizure and decreased GCS on the XX-XX-XXXX and there was no sign of ischaemia or infarct on the repeat CT. She was treated with nimodipine and phenytoin. Phenytoin has now stopped. A repeat CT scan on the XX-XX-XXXX showed a small region of right frontal low density likely due to vasospasm. Mrs. XXXX also developed hospital acquired pneumonia and improved on chloramphenicol.

Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalized idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 – Pneumonia unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 – Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collapse (Q) R56.9 - Unspecified convulsions.

Output: G

Referral letter: Admitted acutely as thrombolysis call with right sided weakness in her arm and leg and speech disturbance. Thrombolysis not performed as diagnosis unclear. CTH and MRI brain normal. Symptoms are ongoing and are thought to be functional. We wonder if there are any further investigations we should be doing to aid a diagnosis or if she needs neurological follow up.

Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalized idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 – Pneumonia unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 – Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collapse (Q) R56.9 - Unspecified convulsions.

Output: M

Referral letter: Dear Dr XXXX, We would be grateful if you could review this 50-year-old Gentleman who you have met previously. He was admitted on the XX-XX-XXXX with severe headache post head injury 2-52 prior. He has had a normal MRA and LP but is struggling with his headache despite regular analgesia, fluids and amitriptyline. Many thanks

Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalized idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 – Pneumonia unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 – Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collapse (Q) R56.9 - Unspecified convulsions.

Output: O

Referral letter: Admitted after fall with loss of consciousness, thought likely secondary to postural hypotension. Reports 2 falls per day. Inflammatory markers raised on admission but settled after treatment for aspiration pneumonia. Normally lives fairly independently at home, no package of care. PMHx AF, CVA XXXX (two tiny sites of hemorrhage involving the right cerebellar hemisphere and corpus callosum), thalassemia trait, portal hypertension Poor oral intake prior to admission, with associated weight loss on background of low BMI previously. BMI 17. Reviewed by speech and language therapy who know her from previously. Found to have severe oropharyngeal dysphagia and advised for NBM and NG feed. Felt to be significantly worse than during previous admissions; cause of deterioration unclear.

Fatigability noted over the course of assessment. CT head and MRI for further CVE demonstrated no evidence of infarct or hemorrhage. Small vessel disease including the pons was noted. Speech is of a nasal quality, with ataxic dysarthria. Awaiting video fluoroscopy Friday, and AChR antibodies. We would be grateful for your opinion as to whether her symptoms could all be explained by previous stroke with small vessel disease of the brainstem, or whether this might represent an alternative etiology such as myasthenia or bulbar palsy. Many thanks.

Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalized idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 – Pneumonia unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 – Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collapse (Q) R56.9 - Unspecified convulsions.

Output: P

Referral letter: Mr XXXX, 72 year-old gentleman, was admitted post-radical chemo-radiotherapy for SCC esophagus with reduced mobility and poor oral intake. On assessment we are concerned that some of these symptoms may be related to possible Parkinsonism.

Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalized idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 – Pneumonia unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 – Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collapse (Q) R56.9 - Unspecified convulsions.

Output:

5.4.2 Data Augmentation

Section 5.1 highlighted the challenges associated with the labor-intensive and expensive process of labelling and anonymizing referral letters, which consequently results in a paucity of publicly available free text based clinical datasets. Through collaboration with neurologists from UHD, a labelled referral letter dataset has been compiled, as depicted in Figure 5.2 and detailed in Table 5.3. However, the size of this dataset is relatively limited for the downstream neural network training, and there is an imbalanced distribution of cases across diseases, potentially affecting its efficacy for subsequent tasks. Drawing inspiration from the methods outlined by ChatAug (Dai et al. 2023), Aplaca (Taori et al. 2023), and ZeroShotDataAug (Ubani et al. 2023), this chapter leveraged the prompted text generation capabilities of ChatGPT for dataset augmentation, demonstrating noteworthy performance improvements in downstream disease classification task.

Given datasets D representing disease names and R representing referral letters, this study employs the few-shot text generation capabilities of ChatGPT to produce an augmented dataset, denoted as R_{aug}. The corresponding pseudocode is presented in Algorithm 2. During this procedure, two parameters are defined: $data_{num}$ and $perturn_{num}$, which respectively control the total number of generated referral letter instances and the quantity of referral letters generated in each iteration. Due to latency in the API response of ChatGPT, empirical tests determined that setting $perturn_{num} = 2$ yields optimal results. The flowchart detailing data augmentation is presented in the left section of Figure 5.3, where the "prompt module" corresponds to the prompt delineated in Algorithm 2.

Algorithm 2 The framework of using ChatGPT to augment referral letter dataset

Input: Disease name set D, Referral letter R which contains the paired referral letter R_{referral} and its corresponding disease name R_{disease}. $R = \{R^i\} = \{[R^i_{\text{referral}}, R^i_{\text{disease}}]\}$ where $i = 1, \ldots, m$ represents the number of diseases

Initialization: Initialize chatgpt-3.5 API key, the generated dataset $R_{\text{aug}} = []$.

Parameters: $data_{num}$ is the total number of generated referral letters, $perturn_{num}$ is the number of generated referral letters per-turn. $Instruct$ = "Based on the following examples, please generate " + str($perturn_{num}$) + " referral letters for patients diagnosed with the disease "

for each $disease_{name}$ in D **do**
    for $turn$ in 0 to $data_{num}/perturn_{num}$ **do**
        rd_{referral} = random(R_{referral}, 2) when $disease_{name} = R_{disease}$
        $prompt = Instruct + disease_{name}$ + "Referral letter 0: " + $rd_{\text{referral}}[0]$ + "Referral letter 1" + $rd_{\text{referral}}[1]$
        $gpt_{resp} = \text{ChatGPT}(prompt)$
        $R_{\text{aug}} = R_{\text{aug}} \cup [gpt_{resp}, disease_{name}]$
    end for
end for

5.4.3 Embedding visualisation and evaluation

This section presents a visualisation of both the augmented dataset and the original collected dataset within a shared embedding space, aiming to discern the distributional proximity of the generated dataset to authentic referral

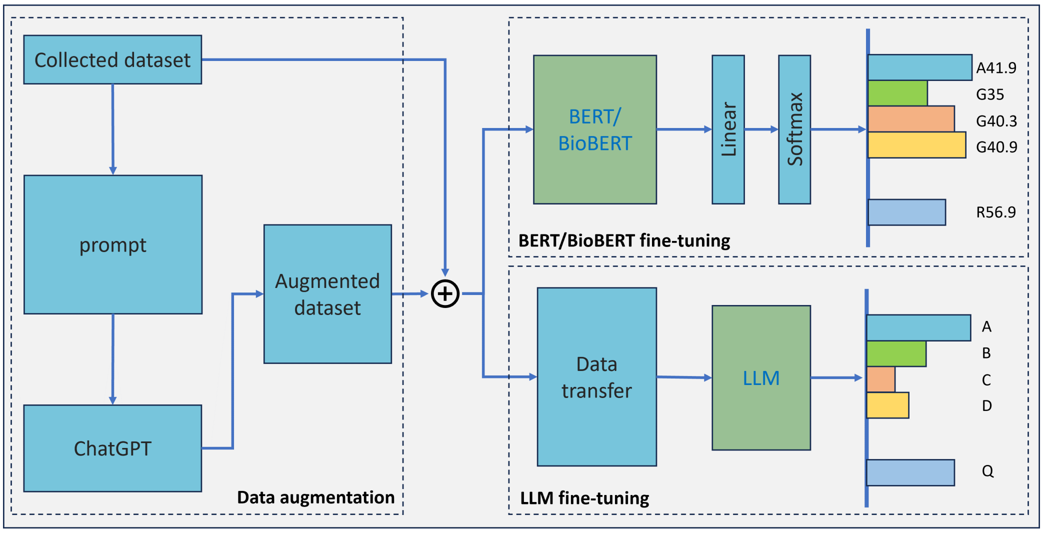


Figure 5.3: Flowchart of data augmentation and downstream fine-tuning.

letters. For embedding the referral letters, the BioBERT (Lee et al. 2020) model is utilised. The embedding of a given referral letter r can be formulated as:

$$Embedding = BioBERT(r) \tag{5.9}$$

where the Embedding corresponds to the representation of the $[CLS]$ token.

Prior to assessing the similarity between a generated referral letter and its corresponding actual referral letter for a specific disease, a formal definition is introduced. Analogous to the description in the Data Augmentation section, the dataset of collected referral letters is denoted as $R = \{R^i\} = \{[R^i_{referral}, R^i_{disease}]\}$ where $i = 1, \ldots, m$ represents the number of diseases. For each $R^i_{referral} = \{r^i_j\}$ where $j = 1 \ldots n_i$ denotes the number of referral letters corresponding to disease i. Similarly, the augmented dataset is represented as $R_{aug} = \{R^i_{aug}\}$ for $i = 1, \ldots, m$, and for each $R^i_{aug} = \{r^i_{aug(j)}\}$ for $j = 1, \ldots, n^i_{aug}$ corresponds to the number of augmented referral letters for disease i. To quantify the disparity and similarity between generated referral letters and genuine test dataset samples, the Kullback–Leibler (KL) divergence $D_{KL}(r^i_{aug(j)}||R^i)$ is utilised, which is shown in the following algorithm.

Detailed computations are provided in Algorithm 3, where S_j^i represents the mean similarity between a generated referral letter and the entire set of collected referral letters for the same disease.

Algorithm 3 Similarity between a generated referral letter and the actual referral letter under the same disease

1: **Definition:** the collected dataset $R^i_{referral} = \{r^i_j\}$ and the generated dataset $R_{\text{aug}} = \{R^i_{\text{aug}}\}$ where $R^i_{\text{aug}} = \{r^i_{\text{aug}(j)}\}$.

2: **Initialization:** Initialize the *BioBERT* model

3: **for** i **in** range(0, m) **do**

4:     **for** j **in** range(0, n_{aug}) **do**

5:         $S_j^i = \frac{D_{KL}(r^i_{aug(j)}||R^i)}{\text{len}(R^i)} = \frac{1}{\text{len}(R^i)} \sum_{k=1}^{\text{len}(R^i)} BioBERT\left(r^i_{\text{aug}(j)}\right) \times$

6:         $[log\left(BioBERT\left(r^i_{\text{aug}(j)}\right)\right) - log\left(BioBERT(r^i_k)\right)]$

7:     **end for**

8: **end for**

5.4.4 Disease classification task

To identify the optimal solution for aiding GPs with referral letters, this section introduces several methods leveraging the augmented dataset. Concurrently, to assess the efficacy of fine-tuning LLMs on free-text-based disease classification, two distinct mechanisms are proposed: the encoder-based PLMs fine-tuning and the LLMs fine-tuning. The flowchart of the two mechanisms can be found in the right section of Figure 5.3.

As previously defined, the augmented dataset can be represented as $R_{aug} = \{R^i_{aug}\} = \{r^i_{aug(j)}\}$. The label of the referral letter, corresponding to its associated disease name, is denoted as $D_{aug} = \{d^i_{aug(j)}\}$ where $i = 1 \ldots m$ signifies the number of diseases and $j = 1 \ldots n^i_{aug}$ indicates the count of referral letters for disease i.

5.4.4.1 Encoder-based PLMs Fine-tuning

BioBERT, pre-trained on a vast corpus of medical text, is further fine-tuned using the augmented dataset. Additionally, given the widespread adoption of the BERT model for text embeddings, the augmented data is used to train it, aiming to identify a more optimised solution for assisting GPs in interpreting referral letters. For given discrete probability distributions d and p with the same support, the primary objective of pre-training encoder-based PLMs is to minimise the cross-entropy loss as defined:

$$L(d,p) = -\sum_{i=1}^{m}\sum_{j=1}^{n^i_{\text{aug}}} d^i_j \cdot log(p^i_j) \tag{5.10}$$

Subsequently, tokens are represented using both token embeddings and positional embeddings, as given by:

$$h^i_j = r^i_{\text{aug}(j)} W_e + W_p \tag{5.11}$$

Where W_e is the token embedding matrix, W_p is the position embedding matrix and h^i_j is the tokenized vector for the input $r^i_{aug(j)}$. The encoder-based PLMs, namely BERT and BioBERT, are then employed to extract the embedding features from the input referral letters:

$$H^i_j = \text{encoder}_{model}(h^i_j) \tag{5.12}$$

then a Linear classifier module is used, coupled with a Softmax activation, to process the feature of the input referral letter $r^i_{aug(j)}$, determining its distribution for possible diseases as follows:

$$p^i_j = \text{Softmax}(\text{Linear}(H^i_j)) \tag{5.13}$$

5.4.4.2 LLM Fine-Tuning

Before fine-tuning the LLMs, a data transfer module, as illustrated in the right section of Figure 5.3 should be introduced. This module converts the augmented dataset into a multiple-choice question-answer (QA) format. In this format, the disease prediction categories serve as choices and are incorporated into the input text. The objective is to predict the correct choice from the provided options. A comprehensive description of this module can be found in Section 5.5.3.3. For each $r^i_{aug(j)}$ in R_{aug}, it comprises K tokens and can be expressed as $r^i_{aug(j)} = \{s_1, s_2, \ldots, s_K\}$. The primary objective when fine-tuning an LLM is to maximise the following likelihood function:

$$L\left(r^i_{\mathrm{aug(j)}}\right) = \sum_{k=1}^{K} \log P\left(s_k \mid s_1, s_2, \ldots, s_{k-1}; \theta\right) \tag{5.14}$$

where θ denotes the trainable parameters of the LLM.

Then the tokens are represented through both token and positional embeddings:

$$h^i_j = r^i_{\mathrm{aug}(j)} W_e + W_p \tag{5.15}$$

Here, W_e and W_p have the same meaning as in the section 5.4.4.1. The subsequent transformation is given by:

$$H^i_j = \mathrm{LLM}(h^i_j) \tag{5.16}$$

It's imperative to highlight that as the size of models escalates, comprehensive fine-tuning becomes impractical on standard consumer hardware. Furthermore, the independent storage and deployment of fine-tuned models for distinct downstream tasks become economically prohibitive, primarily since the fine-tuned model dimensions mirror those of the initial pre-trained model. Parameter-Efficient Fine-Tuning (PEFT) (Mangrulkar et al. 2022) offers a streamlined adaptation of pre-trained LLMs to a plethora of downstream tasks without necessitating a complete parameter fine-tuning, thereby substantially curtailing both temporal and computational overheads. This study adopt the Low-Rank Adaptation (LoRA) technique (Hu et al. 2021), a method acknowledged for its efficacy. This research has indicated that when a pre-trained model undergoes training for a specific subtask, over 90% of the performance achieved by full fine-tuning for that subtask can be realised by adjusting only a limited number of parameters. Upon adaptive fine-tuning, the resultant parameter matrix tends to exhibit a reduced intrinsic dimension. Consider the parameter matrix $W_0 \in \mathbb{R}^{d\times k}$ of a pre-trained model; $\Delta W \in \mathbb{R}^{d\times k}$ signifies the modified parameter matrix post-fine-tuning for a designated task. Given the low intrinsic dimension of ΔW, it can be represented as:

$$W_0 + \Delta W = W_0 + BA \tag{5.17}$$

where B and A are trainable matrices of low rank, and $rank(A), rank(B) \ll \min(d, k)$, thereby enhancing training efficiency and conserving memory space. Thus, the forward pass $h = W_0 h_j^i$ in the LLM can be reformulated as:

$$h = W_0 h_j^i + \Delta W h_j^i = W_0 h_j^i + BA h_j^i \tag{5.18}$$

Finally, the target token is predicted using the Linear weight matrix W_e^T expressed as:

$$s_i = \text{softmax}(H_j^i W_e^T) \tag{5.19}$$

Considering computational constraints and open-source availability, LlaMa-7B and LlaMa-13B are chosen as the LLMs for disease classification. Detailed model configurations will be elucidated in Section 5.5.

5.5 Experiments and Results

This section elucidates the experimental settings in detail, referencing the methodology delineated in Section 5.4. Initially, both direct and multiple-choice QA mechanisms are examined for disease diagnosis using ChatGPT and GPT-4, particularly when provided with a referral letter. Subsequently, an experiment was conducted to investigate data augmentation leveraging the text generation capabilities of ChatGPT. Concurrently, this section visualises the relationships between the augmented referral letters and authentic clinical referral letters collected, as represented in the embedding space. Finally, the parameter configurations adopted for fine-tuning downstream encoder-based PLMs and LLMs using the augmented dataset are detailed.

5.5.1 Hardware Configuration

Our machine operates on an Ubuntu 20.04.4 LTS operating system, equipped with 128GB of RAM and dual NVIDIA RTX A5000 GPUs, each boasting 25GB of GDDR6 memory. The central processing unit (CPU) is based on the x86_64 architecture, specifically the AMD Ryzen Threadripper PRO 3975WX with 32 cores, 64 threads, a base clock speed of 3.5GHz, and a turbo boost frequency of 4.2GHz. Additionally, the system possesses 2TB of solid-state storage.

5.5.2 Baselines Methods

The experimental procedure employed various models to ascertain the optimal language models or processes for aiding GPs in disease diagnosis using free-text referral letters.

ChatGPT: Due to instances where GPT-3's generated outputs might be misleading, offensive, or manifest detrimental sentiments, OpenAI meticulously curated its training dataset. A more refined reinforcement learning from human feedback strategy was applied to GPT-3, leading to the development of ChatGPT. This model possesses 175B parameters and demonstrates enhanced performance in in-context learning tasks. Access to ChatGPT is facilitated through its API. Given the experiment's constraints on input tokens, this section opted for the gpt-3.5-turbo API with a maximum 4K context. The associated costs are \$0.0015 per 1K input tokens and \$0.002 per 1K output tokens.

GPT-4: GPT-4, similar to ChatGPT, is grounded in the GPT architecture but boasts a significantly larger structural scale, estimated at 100T parameters. As per GPT-4's report (OpenAI 2023), it showcases human-equivalent proficiency across diverse professional and academic benchmarks, including achieving a score within the top 10% of participants in a simulated bar exam. Similar to its predecessor, GPT-4 can be accessed via an API. For research purposes, this chapter employed the GPT-4 API with a maximum 8K context, incurring costs of \$0.03 per 1K input tokens and \$0.06 per 1K output tokens.

BERT: The BERT (Bidirectional Encoder Representations from Transformers) model, introduced by Google, has significantly transformed the landscape of NLP. Contrary to the autoregressive architecture of GPT, BERT is designed to predict masked words within a given sentence, thereby offering bi-directional capabilities. In this research, the "bert-base-cased" model is utilised, which is characterised by a 12-layer architecture, 768 hidden units, 12 attention heads, and a total of 110 million parameters.

BioBERT: BioBERT serves as a domain-specific adaptation of BERT tailored for biomedical texts. This model undergoes pre-training on expansive biomedical corpora. Consequently, for tasks such as biomedical information retrieval, biomedical named entity recognition, and related challenges, BioBERT may exhibit superior performance compared to the standard BERT due to its nuanced understanding of domain-specific terminologies and contexts. The "biobert-base-cased-v1.2" model is selected in the research to embed referral letters. This model also possesses a 12-layer structure, 768 hidden units, 12 attention heads, and a complement of 110 million parameters.

LlaMa: The LlaMa is a publicly available LLM developed by Meta AI. Notably, LlaMa is trained on an expansive dataset comprising 1 trillion tokens sourced from open-access data repositories. Empirical evaluations reveal that LlaMa offers competitive performance relative to the considerably more parametric GPT-3, which boasts 175 billion parameters across various NLP benchmarks. Considering the capabilities of the hardware configurations, the LlaMa-7B and LlaMa-13B models have been selected for the designated downstream tasks.

LoRA: The LoRA approach presents an efficacious adaptation strategy for LLMs. It would not induce additional inference latency or truncate the input sequence length while preserving superior model quality during fine-tuning processes. Throughout the training phase, LoRA freezes the pre-trained model weights as static and introduces trainable rank decomposition matrices to each layer within the Transformer architecture. This methodology substantially curtails the quantity of trainable parameters necessary for downstream applications.

The LLaMA-7B architecture is composed of 32 transformer blocks accompanied by three fully connected output layers. Storing its checkpoint necessitates a minimum storage capacity of 23 GB. Full fine-tuning mandates the use of at least two GPUs, each possessing a minimum memory of 30 GB, and entails the implementation of full sharding for weight allocation. As an alternative, one can deploy four GPUs, each demanding up to 22 GB of memory, an option that is incompatible with our current equipment. Experimental results indicate that with a rank of 8 for LoRA, the trainable parameters during the fine-tuning phase can be minimised to 8 MB. Consequently, this study employed the LoRA adaptation strategy, which significantly diminished the computational overhead without compromising training precision.

5.5.3 Experiment design and results

Given the primary objective of this study, which is to identify an optimal solution for assisting GPs in decision-making using referral letters and to validate the efficacy of LLMs in referral letter-based disease classification, this section employed the methods delineated in Section 5.4. Specifically, in-context learning diagnostics are designed for ChatGPT and GPT-4 based on the collected 439 referral letters, as statistically detailed in Table 5.2, incorporated data augmentation strategies, and conducted fine-tuning on PLMs. Comprehensive details of the experiments, along with the corresponding results, are presented in the subsequent section.

5.5.3.1 ChatGPT and GPT-4

Direct diagnosis on ChatGPT and GPT-4

Drawing upon the prompt examples presented in Table 5.4, all the real collected clinical referral letters are evaluated using both ChatGPT and GPT-4, employing zero-shot and one-shot prompting methodologies. As outlined in Section 5.4.1.1, within the encoder-based PLM approach to medical text classification, symptom information plays a pivotal role in determining prediction accuracy. Consequently, the symptom data is incorporated into the prompt text to ascertain its impact on the outcomes. The diagnostic accuracy, utilising in-context learning, is delineated in Table 5.6. Herein, the metrics zero-shot-symp and one-shot-symp represent the prediction accuracy associated with symptom-inclusive prompts.

Table 5.6: The accuracy of direct diagnosis on ChatGPT and GPT-4

Model	**zero-shot**	**zero-shot-symp**	**one-shot**	**one-shot-symp**
ChatGPT	0.118	0.120	0.333	0.344
GPT-4	0.203	0.228	0.294	0.296

By examining the data presented in Table 5.6, the subsequent conclusions can be draw: 1) Compared with the zero-shot for the ChatGPT on the disease prediction, there is a marked improvement in one-shot accuracy by 21.5%. When compared with the zero-shot-symp prompt for ChatGPT, the one-shot-symp accuracy witnesses an increase by 22.4%. Similarly from the GPT-4 results, the one-shot accuracy saw an enhancement of 9.1%, while the prediction accuracy for one-shot-symp rose by 6.8% than the corresponding zero-shot prompt. These increments corroborate the empirical observations indicating that the integration of samples into the prompt text augments the performance of GPT-based generative models. Nonetheless, the diagnostic prediction accuracy remains unsatisfactorily low for genuine clinical scenes, rendering it unsuitable for general practitioners.

2) To evaluate the efficacy of integrating symptoms into the prompt text, this study contrasted the accuracies of zero-shot with zero-shot-symp, and one-shot with one-shot-symp for both ChatGPT and GPT-4. The findings elucidate that the inclusion of symptoms within the referral letter during the zero-shot phase enhances the accuracy for ChatGPT by 0.2%, while GPT-4 experiences a more improvement of 2.5%. On the other hand, introducing symptoms post the referral letter in the one-shot phase results in an accuracy boost for ChatGPT by 1.1%, whereas GPT-4's accuracy increment is a mere 0.2%. These results underscore that directly incorporating symptoms into the referral letters in the prompt text can refine prediction accuracy for

both zero-shot and one-shot learning, albeit the enhancements are relatively minimal.

3) Assessing each data format and comparing the prediction accuracies between ChatGPT and GPT-4, it is evident that for the zero-shot and zero-shot-symp prompt data, GPT-4's prediction accuracy surpasses ChatGPT by an approximate margin of 10%. Conversely, with the one-shot and one-shot-symp data, ChatGPT demonstrates superior performance, leading GPT-4 by 4-5%. The findings suggest that, while the parameter scale of GPT-4 exceeds that of ChatGPT, GPT-4's performance may not necessarily surpass ChatGPT in certain domain-specific tasks.

Multiple-choice question and answering

The accuracy of ChatGPT and GPT-4, when employed direct diagnosis, leaves much to be desired. To further explore the feasibility of applying ChatGPT or GPT-4 to real-world diagnostic scenes and drawing inspiration from the multiple-choice QA paradigm of LLMs, the GP disease diagnosis process is restructured into a multiple-choice QA format, with disease categories serving as the answer options. Adhering to the methodology delineated in Section 5.4, this study utilised the prompt detailed in Table 5.5 for ChatGPT and GPT-4 interrogation. The resulting prediction accuracy is documented in Table 5.7.

Table 5.7: Accuracy of multiple-choice QA disease prediction

Model	**zero-shot**	**one-shot**	**Few-shot**
ChatGPT	0.141	0.355	0.405
GPT-4	0.100	0.394	0.544

As depicted in Table 5.7 and in comparison, to the direct diagnosis results from Table 5.6, the use of the proposed multiple-choice QA methods demonstrates an improvement of 2.3% in zero-shot prompt for ChatGPT, while indicating a decrease of 10.3% for GPT-4. The one-shot prompt accuracy, when utilising the proposed method, improves by 2.2% for ChatGPT and 10% for GPT-4. Additionally, Table 5.7 reveals that the few-shot prompt considerably enhances the prediction accuracy of both GPT models. However, even with GPT-4's accuracy peak of 54.4%, this performance is risky for directly used in the real clinical scenes.

5.5.3.2 Data Augmentation

Utilising Algorithm 1 for data augmentation, as described in Section 5.4.2, this study generated datasets of 100, 200, and 500 cases for each respective disease, and then those three augmented datasets are connected together as an 800 cases dataset. Then BioBERT is employed as the encoder model to derive the embedding representations of each referral letter. Figure 5.4 visualises both the collected referral letters and the generated referral letters of 800 cases within the same embedding space. In this figure, the blue points correspond to the embeddings of the actual collected clinical referral letters, while the orange points denote the embeddings of the augmented referral letters. It is noteworthy that, although a few orange points are located at the periphery, away from the centre of the real referral letter embeddings, the majority of the generated data is concentrated in the central region of the embedding space. This suggests that their representational meaning is closely aligned, affirming that the dataset generated via Algorithm 1 serves as a suitable augmentation of the collected dataset.

In addition,Algorithm 3 quantifies the similarity between the generated and collected referral letters, depicted in Figure 5.5 as a boxplot.

Here, the x-axis lists the 17 diseases studied, and the y-axis shows KL divergence between generated and actual letters. Figure 5.5 suggests that the KL divergence for most generated letters is lower than 0.05, indicating the generated referral letters are aligned with actual ones. This similarity promotes favourable outcomes in data-intensive deep learning models. Besides, Figure 5.5 also shows outliers for each disease, which can potentially enhance model generalizability. Nevertheless, a KL divergence concentrated around 0.05 implies a close correspondence between the generated referral letters and their respective authentic referral letters, as well as a proximity within

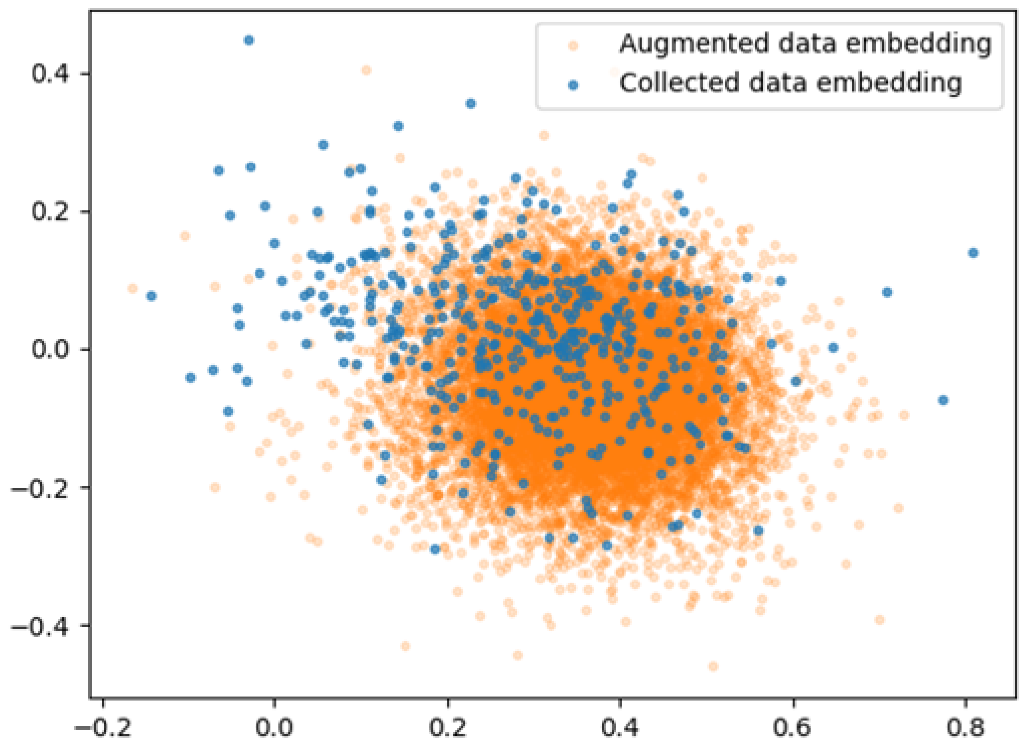


Figure 5.4: Visualization of referral letters in the embedding space

the generated referral letters themselves. This proximity raises the potential concern of overfitting, thereby impacting the formulation of downstream fine-tuning experimental configurations.

5.5.3.3 Fine-tuning on the PLMs

This section has three primary objectives. Firstly, this section aims to develop an accurate solution that aids GPs in diagnostic decision-making using referral letters. Secondly, the performance of fine-tuning LLM for the disease prediction task is validated. Lastly, this section explores the extent to which the augmented dataset enhances performance in designed downstream tasks. To this end, experiments on two downstream tasks are conducted: 1) supervised fine-tuning using traditional text classification models such as BERT and BioBERT; and 2) multiple-choice-based QA predictions leveraging the

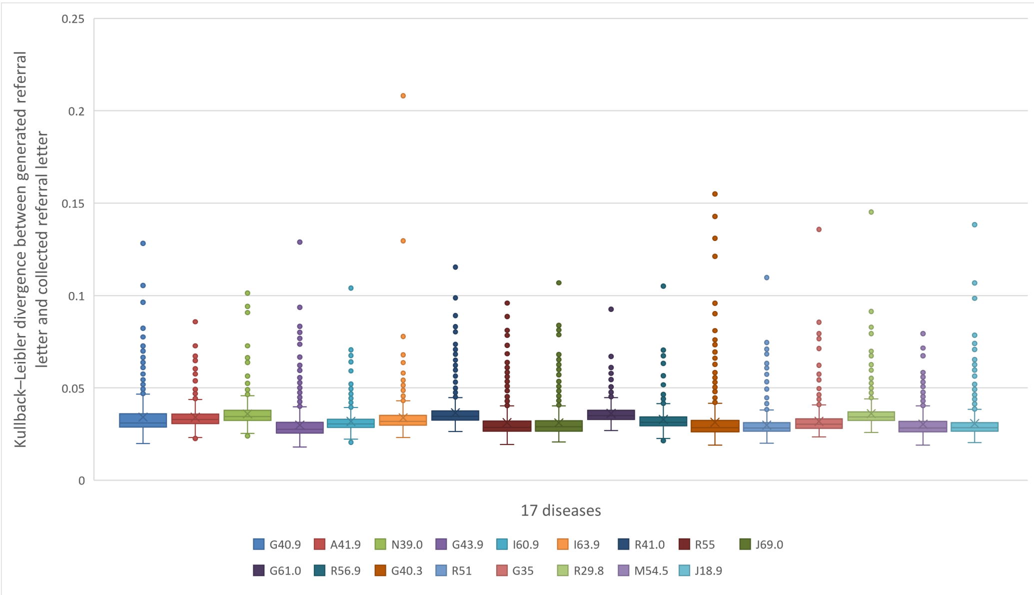


Figure 5.5: Kullback–Leibler divergence between generated referral letters and collected referral letters

LLM. To achieve the objectives of this chapter, the augmented datasets have sample sizes of 100 cases, 200 cases, and 500 cases for each disease, cumulatively amounting to 800 cases for each disease. These datasets were utilised for downstream disease prediction employing BERT, BioBERT, LlaMa-7B, and LlaMa-13B. It is noteworthy that the current hardware supports only LlaMa-7B. Thus, this study rented four A100 units for LlaMa-13B training. Due to budgetary constraints, only the 500-case dataset was evaluated using LlaMa-13B.

To comprehensively assess the effectiveness of both the augmented and collected data in downstream fine-tuning tasks and investigate the influence of the potential overfitting arising from a highly concentrated augmentation dataset, four experimental setups are devised:

- mixed: This setup combines both the generated referral letters and the actual collected referral letters. Both data types are randomly distributed across the training, testing, and evaluation datasets. This design seeks to assess model performance on the "mixed dataset" holistically, setting aside concerns of overfitting.

- gen-train-collect-test: This approach utilises the entire generated dataset for training the model, while the real collected dataset serves the purpose of evaluation. Contrary to the mixed setup, this configuration attempts to mitigate overfitting to some extent. The objective is to determine the influence of the augmented dataset on downstream fine-tuning outcomes for both encoder- and decoder-based PLMs.

- mixed-train-collect-test: In this configuration, the training set consists of both generated and collected data, while the testing set exclusively uses real collected data. While it shares some similarities with the "gen-train-collect-test" paradigm in its efforts to circumvent overfitting, it integrates real collected data into the training set to bolster authentic knowledge transfer to the model. This could potentially enhance accuracy in disease prediction. Therefore, the outcomes from this paradigm are mostly cared in the experiment and can serve as an indicator of the quality of the generated dataset as well as the performance of the fine-tuned model.

- generation-only: This experimental design uses only the generated data for training, testing, and evaluation. Intended as a control setup, its design is aimed at quantifying the extent of overfitting attributable to data redundancy.

For fine-tuning on the encoder-based PLMs, BERT and BioBERT are used as training, and they share the same training configuration. The model is configured with a batch size of 8, a dropout rate of $p = 0.5$, and is optimised using the Adam algorithm (Kingma and Ba 2014). The learning rate is set to $1e - 6$, with training conducted over 100 epochs. The dataset is partitioned into training, testing, and evaluation sets at a ratio of $3 : 1 : 1$, respectively. Detailed information can be found in Table 5.8.

For fine-tuning on the decoder-based LLM, the block sizes for LlaMa-7B and LlaMa-13B are set at 512. The batch size is fixed at 1, with a LoRa

Table 5.8: Train/test/eval data scale for encoder-based model

Experiments	100 cases			200 cases			500 cases			800 cases		
	train	test	eval	train	test	eval	train	test	eval	train	test	eval
mixed	1222	407	408	2212	737	738	5087	1695	1697	6036	2012	2012
mixed-train-collect-test	1817	220	220	3467	220	220	8259	220	220	13105	220	220
generation-only	958	319	321	1949	650	652	4824	1608	1608	-	-	-
gen-train-collect-test	1451	293	293	3102	294	294	7893	293	293	9474	293	293

rate r=120. The total number of parameters is 6, 801, 330, 176, of which 62,914,560 are trainable. This yields a trainable parameter percentage of 92.5%. This study employed a learning rate of 1e-4 and utilised the Adam optimisation algorithm. For each experiment, the train-to-test data ratio is configured at 3:1. Comprehensive details are provided in Table 5.9.

Table 5.9: train/test data scale for LLMs

Experiment	100 cases		200 cases		500 cases		800 cases	
	train	test	train	test	train	test	train	test
mixed	1520	539	2738	949	6338	2141	10655	3384
mixed-train-collect-test	1940	119	3568	119	8360	119	13920	119
generation-only	1200	420	2418	830	6018	2022	9621	3265
gen-train-collect-test	1620	439	3248	439	8040	439	13600	439

Finally, this study has consolidated the training results of both BERT and LlaMa and presented them in Table 5.10. Upon comparing and analysing the prediction accuracy as presented in Table 5.10, the following observations are deduced:

Overall comparison:

Table 5.10: Prediction accuracy of models on designed setups

Models	**Experiment setups**	**100 cases**	**200 cases**	**500 cases**	**800 cases**
LlaMa-7B	mixed	0.459	0.717	0.881	0.905
	mixed-train-collect-test	0.034	0.336	0.664	0.723
	generation-only	0.334	0.673	0.871	0.905
	gen-train-collect-test	0.200	0.228	0.524	0.592
BERT	mixed	0.902	0.963	0.974	0.981
	mixed-train-collect-test	0.800	0.864	0.923	0.977
	generation-only	0.918	0.957	0.977	-
	gen-train-collect-test	0.727	0.785	0.874	0.877
BioBERT	mixed	0.885	0.947	0.979	0.978
	mixed-train-collect-test	0.782	0.841	0.968	0.973
	gen-train-collect-test	0.710	0.802	0.894	0.932
LlaMa-13B	mixed			0.873	
	mixed-train-collect-test			0.621	
	generation-only			0.864	
	gen-train-collect-test			0.526	

For each model, when trained using datasets of 100 cases, 200 cases, 500 cases, and 800 cases under each experimental setup, it becomes evident that the incorporation of generated data significantly enhances prediction accuracy. The trend indicates that as the data scales from 100 cases for each disease to 500 cases, there is a substantial improvement in predictive accuracy. Conversely, when this scale increases from 500 to 800 cases, the rate of improvement decelerates. To elucidate, during the LlaMa-7B training under the mixed-train-collect-test experiment, the accuracy surged by 0.63 when the dataset increased from 100 to 500 cases. However, in the transition from 500 to 800 cases, the accuracy increased by only 0.059. This enhancement suggests that data augmentation techniques exert a beneficial effect on downstream disease classification tasks, and the resulting augmented dataset exhibits a substantial quantity.

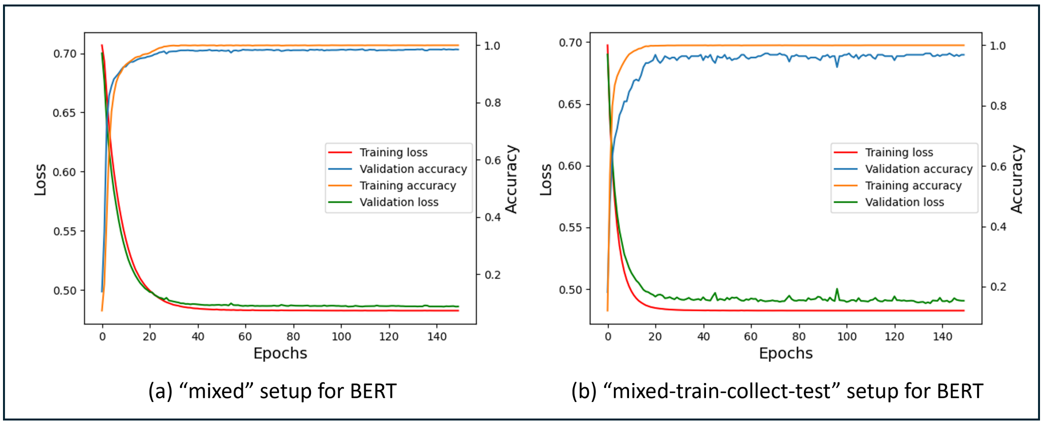
In fields grappling with data sparsity, employing generation-based LLMs such as ChatGPT for data augmentation proves effective and merits exploration. Nonetheless, it is pivotal to acknowledge that the benefits of generated data exhibit diminishing returns. Beyond a certain threshold of generated data volume, the incremental improvements in downstream classification accuracy become relatively marginal.

Comparison of four experiment setups

To mitigate concerns related to overfitting and bias arising from the augmented dataset, this section establishes four experimental configurations encompassing various combinations of generated data and actual referral letters. Table 5.10 presents the predictive accuracy of these setups across different models, revealing a consistent trend. Specifically, when the training dataset for each disease exceeds 100 cases, the predictive accuracy follows the order: $mixed > generation - only > mixed - train - collect - test > gen - train - collect - test$. This observed trend aligns with the design principles elucidated in Section 5.5.3.3.

Given that both "mixed" and "mixed-train-collect-test" setups incorporate actual referral letters, their performance warrants focused attention. Consequently, Figure 5.6 illustrates the performance of these experimental setups on the BERT model, which demonstrated the highest classification accuracy among the four setups in the scenario involving 800 cases. Figure 5.6 facilitates a comprehensive understanding of the model and dataset efficacy during the training phase and subsequent validation. The left vertical axis portrays the loss, with the red line indicating training loss and the green line representing validation loss. On the right vertical axis, accuracy is depicted, where the blue line signifies validation accuracy and the orange line denotes training accuracy. The horizontal axis corresponds to the number of training epochs.

Figure 5.6.a illustrates the performance of the mixed setup, depicting a noteworthy issue from epochs 0 to 20. During this period, the validation loss consistently surpasses the training loss, while the validation accuracy concurrently outperforms the training accuracy. This anomaly is indicative of



(a) “mixed” setup for BERT

(b) “mixed-train-collect-test” setup for BERT

Figure 5.6: Performance of the experiment setups for BERT. (a) “mixed” experiment setups for BERT; (b) “mixed-train-collect-test” experiment setups for BERT.

a potential “data leakage” scenario, wherein certain generated data exhibit notable similarities and are allocated to both the training and validation datasets. Consequently, it is imperative to confine the inclusion of generated data solely to the training dataset for this disease classification task, mitigating the risk of “data leakage”. Conversely, the performance of the “mixed-train-collect-test” setup, as depicted in Figure 5.6.b, exhibits commendable performance, as evidenced by the coherent trends observed in loss and accuracy metrics for both training and validation datasets. This finding suggests that the employed “mixed-train-collect-test” methodology adheres to a reasonable dispatch principle. With this principle, the generated dataset exhibits noteworthy quality and demonstrates a significant improvement in disease classification accuracy with an increase in the size of the generated training dataset. Consequently, the “mixed-train-collect-test” setup proves valuable for assessing the performance of various models in subsequent discussions related to disease classification.

Comparison of BERT and BioBERT

Excluding the “mixed” and “generation-only” configurations due to concerns of overfitting, analysis indicated that both BioBERT and BERT achieve similar prediction accuracy levels in the “gen-train-collect-test” and “mixed-train-collect-test" scenes, particularly when considering datasets with fewer than 100 cases. However, as the dataset size expands, BioBERT consistently surpasses BERT in performance. This superior performance can be attributed to the specialised training of BioBERT on extensive medical datasets, which imparts more domain-specific knowledge compared to BERT's broader, generalist training. Notably, within the "mixed-train-collect-test" configuration, BERT achieves a prediction accuracy of 0.977, marking the highest recorded result in this study.

Comparison between BERT/BioBERT and LlaMa-7B

In evaluating the results of the "mixed-train-collect-test" and "gen-train-collect-test" predictions, this study observed that encoder-based PLMs such as BERT and BioBERT consistently outperformed decoder-based LLMs in free-text classification tasks. This superiority can be attributed to the fact that encoder-based models generate a probability distribution over a fixed category. Conversely, the decoder-based model's output, in terms of length, format, and other parameters, remains difficult to control. Even when disease categories are presented as options within the prompt text, managing model outputs remains a challenge. Furthermore, the hardware and training expenses associated with BERT/BioBERT are substantially less than those of LLMs. Therefore, for text classification tasks involving fixed categories, encoder-based PLMs appear to be a more efficient choice over LLMs.

Comparison of LlaMa-7B with LlaMa-13B

Due to the dual A5000 graphic cards unable to execute the LlaMa-13B model. Consequently, this study secured four A100 GPUs and fine-tuned the model using a dataset of 500 cases. The results indicate that the prediction accuracy of LlaMa-13B closely mirrors that of LlaMa-7B, with minor deviations of no more than 0.2 in any of the four experimental setups.

Comparison of LlaMa-7B with ChatGPT and GPT-4

Table 5.7 demonstrates that, among ChatGPT and GPT-4, the highest predictive accuracy is achieved through the few-shot learning approach of ChatGPT-4, resulting in an accuracy of 0.544. Disregarding the potential impact of overfitting, the results of LlaMa-7B in both the "mixed-train-collect-test" and "gen-train-collect-test" experiments were 0.723 and 0.592, respectively. These scores surpass those of GPT-4. Given the costs associated with the GPT-4 API and concerns related to patient data privacy, open-source LLM fine-tuning may present a more suitable alternative for downstream tasks reliant on patient data.

5.6 Limitations

The size of the referral letters collected from UHD neurologists is limited, encompassing only 17 diseases for classification. For broader application of the proposed pathway across the entire medical system, it is imperative to incorporate referral letters from all disease categories. Additionally, due to hardware and budgetary constraints, the performance of fine-tuning LLMs with larger parameters remains untested. However, encoder-based PLMs have achieved satisfactory performance without necessitating enhanced hardware capabilities.

5.7 Summary

This study aims to ascertain an optimal strategy to assist GPs in decision-making by leveraging referral letters and to assess the capability of LLMs in referral letter-based disease classification. The in-context learning diagnostics for ChatGPT and GPT-4 are designed, utilising a dataset comprising 439 real anonymized referral letters. These diagnostics encompassed direct diagnosis approaches and multiple-choice question answering. The experiment results indicated that GPT-4, when deployed in a few-shot learning setting, yielded superior predictive accuracy. However, this accuracy is not a satisfactory performance for real clinical scenarios. Considering the limited scale of the training dataset in the medical domain, this study utilised the text-generation capabilities of ChatGPT to augment the collected referral letters. This augmented dataset was then used to fine-tune encoder-based PLMs, including BERT and BioBERT, as well as LLMs such as LlaMa-7B and LlaMa-13B. The investigations revealed that while the augmented dataset had a tendency to cause overfitting, meticulous experimental design could counteract this issue, leading to a trustworthy performance in downstream disease classification accuracy. Furthermore, analyses consistently demonstrated that encoder-based PLMs, particularly BERT and BioBERT, outperformed decoder-based LLMs in free-text classification tasks, but the accuracy of LlaMa-7B on the 800 cases "mixed-train-collect-test" setups is better than the few-shot learning on the GPT-4. Notably, BioBERT achieved a classification accuracy of 0.973. Consequently, the findings suggest a promising pathway: utilising ChatGPT for data augmentation, then training encoder-based PLMs, and eventually deploying the trained models to assist GPs in decision-making.

Chapter 6

Conclusion and Future Work

In this chapter, the thesis concludes by encapsulating the primary findings and contributions of the research. Additionally, potential avenues for future work, aimed at addressing the present limitations, are explored and discussed.

6.1 Conclusion

The scarcity of medical resources, accentuated by the COVID-19 pandemic, has exerted pressure on every tier of national health systems globally. The advent of artificial intelligence and big data learning, especially in the realm of medical diagnosis, heralds potential mitigations for these pressures. This thesis introduced methods and solutions to alleviate strain on the health system from two fronts. Firstly, this study suggests the deployment of automated diagnostic tools for patients, thereby facilitating primary diagnostic services and concurrently diminishing demands on the medical system. Secondly, the development of diagnostic support tools for General Practitioners is explored, aiming to bolster their diagnostic proficiency and efficiency, which, in turn, elevates the overarching effectiveness of the medical system.

Firstly, to address **research question Q1** delineated in Chapter 1, this thesis investigates the pipelines and functionalities of a dialogue-based diagnostic system for patients, which encompasses several models including data collection, processing, graph building, and the design of both diagnostic policies and the dialogue system. Given the sparse availability of disease-symptom data, a disease-symptom graph was formulated, guided by the richly detailed descriptions of common diseases and their associated symptoms provided on the NHS website. Furthermore, leveraging the constructed disease-symptom graph, a knowledge-based binary-choice diagnostic policy was developed and integrated into the system, supported by subsequent examples demonstrated via a web application. Comprehensive details of this study have been discussed in Chapter 3.

Secondly, to address **research question Q2 and Q3** delineated in Chapter 1, in a pursuit to enhance the efficiency and accuracy of GPs' decision-making, particularly through training referral letters with AI models, a collaborative effort was initiated with neurologists at the UHD. A collection of referral letters and their corresponding complaint texts were amassed for four specified neural diseases, which include epilepsy-recurrent seizures (G40), headache (R51), dorsalgia (M54), and cerebral infarction (I63). Consequently, a hybrid architecture was introduced with the aim of aiding GPs in improving the accuracy of primary diagnoses. Comparative analyses utilising the neurology dataset underscored that the proposed model boasts commendable classification accuracy. Furthermore, two data augmentation methods were proposed—namely, the Symptom Dot Separating Method and the Complaint-Symptoms Integration Method—which have been empirically validated to enhance the predictive accuracy of neurology disease classifications. Additionally, leveraging the pipelines from a previously investigated dialogue-based diagnosis web application for patients, it was re-implemented using the proposed hybrid architecture to develop an AI diagnosis assistant web application. This tool is designed to proficiently aid GPs in making a primary diagnosis through interactive conversation and text engagement. Detailed discussions and findings pertinent to this study are elucidated in Chapter 4.

Finally, to address **research question Q4** delineated in Chapter 1, and as the LLMs have superior performance on text-based tasks, this thesis assesses the capability of LLMs in referral letter-based disease classification and at the same time ascertain an optimal strategy to assist GPs in decision-making by leveraging referral letters. Testing on a dataset comprising 439 real anonymized referral letters, it can be found that the accuracy of directly using LLMs for disease diagnosis does not satisfy performance for real clinical scenarios. Considering the limited scale of the training dataset in the medical domain, the text-generation capabilities of ChatGPT are utilised to augment the collected referral letters. This augmented dataset was then fine-tuned using encoder-based PLMs, including BERT and BioBERT, as well as LLMs such as LlaMa-7B and LlaMa-13B. The investigations revealed that while the augmented dataset had a tendency to cause over-fitting, meticulous experimental design could counteract this issue, leading to a trustable performance in downstream disease classification accuracy. Furthermore, the analyses consistently demonstrated that encoder-based PLMs, particularly BERT and BioBERT, outperformed decoder-based LLMs in free-text classification tasks. Consequently, our findings suggest a promising pathway: utilising ChatGPT for data augmentation, then training encoder-based PLMs, and eventually deploying the trained models to assist GPs in decision-making. The details of this work have been discussed in Chapter 5.

6.2 Limitations and Future Work

This section discusses the inherent limitations and future work of technical methods used for text classification in the medical field, starting from four key aspects, including: data scarcity, vector representation ability, LLMs for diagnosis, and interpretability of the model.

6.2.1 Data scarcity

The lack of genuine, high-quality training data has always been a problem for methods based on deep learning, which restricts the development of artificial intelligence in the medical field. For example, in the work in Chapter 3, without real clinical training data, a disease symptom graph is constructed and a binary decision strategy is proposed. However, if sufficient training data are available, the weights of each node on the graph will be trained, which would be more effective and convincing than using the frequency of symptom occurrence on the disease-symptoms graph as weights. Moreover, due to the lack of labelled real case recommendations, although data augmentation methods were proposed in Chapter 4, which show superior performance in disease classification tasks, they are limited by the size and scale of the training data. Using the proposed hybrid architecture can only predict 4 types of diseases, while the work in Chapter 5 can predict 17 types of diseases. This results in the proposed method being limited in its development and promotion in actual clinical evaluation and application.

The limited availability of medical datasets primarily stems from multiple challenges, including: 1) Adherence to strict privacy and ethical guidelines, which safeguard sensitive personal information and necessitate patient consent. 2) Technological and quality assurance hurdles, ensuring accuracy, consistency, and interoperability across diverse healthcare data formats and standards. 3) Regulatory barriers and the significant resources required for the management, maintenance, and sharing of these datasets, while also ensuring their scientific validity and representativeness. 4) The inherent monetary and competitive value of medical data, which, due to proprietary interests, often restricts public access to organisational datasets. Consequently, reconciling progress in medical research with ethical, legal, and quality assurance challenges warrants innovative strategies, such as the development of **synthetic datasets and the establishment of secure, compliant data-sharing frameworks**.

6.2.2 Vector representation ability

The facility of vector representation, frequently referred to as embeddings in AI models, facilitates the transformation of words, phrases, and other entities into fixed-size vectors within a continuous space, thereby enabling subsequent processing by neural network models. The capability of embeddings is paramount for downstream tasks, and a model that can accurately map textual information into vector space while minimising semantic loss across diverse contexts can be considered to possess an exemplary embedding architecture. Leveraging the robust self-attention mechanism within the

Transformer architecture, models such as BERT and its derivatives, including BioBERT (Lee et al. 2020), SciBERT (Beltagy et al. 2019), and ClinicalBERT (Huang et al. 2019), have been routinely utilized for text representation and have demonstrated superior performance in downstream tasks, as evidenced by the work in Chapter 4 and Chapter 5. Nonetheless, BERT's capacity to represent input sentences diminishes when handling inputs exceeding 512 tokens. Though models like DocBERT (Adhikari et al. 2019), Reformer (Kitaev et al. 2020), LongFormer (Beltagy et al. 2020), Recurrence over BERT (RoBERT), and Transformer over BERT (ToBERT) (Pappagari et al. 2019) have augmented the embedding of protracted sentences, BERT-based embeddings yet seek enhancement when compared with expansive models like GPT, which can accommodate inputs of up to 32,000 tokens.

In the work of Ostendorff et al., an improved BERT architecture was proposed (Ostendorff et al. 2019) for combining text representation with other types of information, namely metadata and knowledge graph embeddings. Comparing these enhanced embeddings with the regular BERT embeddings used for classification tasks showed performance improvements. The work in Chapter 4 also demonstrated that merging static embedding representations and BERT can enhance the performance of downstream tasks. Therefore, to improve performance or address the differences between the source and target domains, it is necessary to equip language models with more domain-specific knowledge (Liu et al. 2020b). Previous research (Xu et al. 2016, Wu et al. 2016, Xie et al. 2016, Gao et al. 2020) mainly utilised deep learning techniques to encode entities' descriptions, reference sentences, or other text information into the knowledge graph embedding space to bridge the gap between language and knowledge. However, the performance of this method is constrained by the size and quality of the knowledge graph, making it difficult to update and scale. Recently popular models like ChatGPT, GPT-4, Llama, and other LLMs perform excellently on text-based general knowledge tasks but do not perform satisfactorily in specific domains, such as medicine.

However, large-scale pre-trained language models (such as GPT-3, GPT-4, BERT, Roberta, and Llama) are pre-trained for unsupervised tasks using a very large corpus containing general world knowledge. They can be further fine-tuned using additional labelled data across various downstream tasks and applications. Unlike research that encodes text information into knowledge graphs, these pre-trained large language models implicitly learn and explore representations in KG embedding space, and structured knowledge can also be explicitly injected into the pre-trained models. Therefore, the process of knowledge integration can be more interpretable and rational, and can be viewed as a future direction for domain tasks.

6.2.3 LLMs for diagnosis

The results in Chapter 5 show that LLM can be used for data augmentation, but the fine-tuning effect of LLM is not as good as the lower hardware requirement BERT and Biobert. However, LLM is currently moving towards knowledge domain specialisation, model miniaturisation, and low-cost training. If a high-quality, professional medical domain LLM can be proposed, coupled with LLM's excellent zero/one/few-shot performance, it would definitely be very good news for medical diagnosis.

6.2.4 Interpretability of the model

Artificial Intelligence models with high interpretability will play a crucial role in the field of medical diagnostics. The interpretability of AI models can seamlessly intertwine with trust and practical functionality within a healthcare environment. Establishing trust between medical professionals and patients largely depends on the model's ability to transparently convey its diagnostic reasoning. This transparent capability not only nurtures trust between doctors and patients but also allows physicians to verify pathological reasoning at various stages, utilising their own knowledge reserves, and make professional clinical decisions. Furthermore, the explanations provided by AI contribute to creating a safe patient environment, enabling clinical doctors to mitigate risks and devise personalised intervention measures through discernible insights. Additionally, a clear and interpretable AI decision-making process has educational attributes and can serve as a learning tool for medical professionals. It can also promote a broader application of AI tools and integrate them into traditional healthcare infrastructure by aligning AI outcomes with established clinical reasoning and practices.

From a technological standpoint, the Chain of Thought (COT) (Wei et al. 2022) and Automatic Chain of Thought (Auto-CoT) (Zhang et al. 2022b) methodologies demonstrate potent efficacy in directing LLMs toward adopting analogous reasoning pathways when responding to prompts, accomplished through the provision of prompt exemplars accompanied by methodical reasoning processes. Empirical investigations substantiate that the implementation of this technique can precipitate a noteworthy augmentation in model performance, particularly on tasks necessitating intricate reasoning capabilities. Consequently, when applied within the context of medical decision-making, this approach holds significant promise for addressing the prevailing challenges related to the interpretability of medical models.

6.2.5 Future work

As elucidated in Chapter 1, the foundation of this thesis is firmly anchored in the Intel-PA project. This project is a teamwork and my role within this project entailed translating the pragmatic requirements provided by the doctors into an academic framework and subsequently proposing academically oriented solutions. Owing to the development schedule and the intricate policy of the collaborated hospital, the evaluation of the developed web applications—for both patients and GPs—in genuine clinical contexts is presently under development by my colleagues. Upon completion of this process, it is anticipated that the progress and practicality of the work will be elevated to an enhanced level.

References

Adhikari, A., Ram, A., Tang, R. and Lin, J., 2019. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.

Akrout, M., Gyepesi, B., Holló, P., Poór, A., Kincső, B., Solis, S., Cirone, K., Kawahara, J., Slade, D., Abid, L. et al., 2023. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. *arXiv preprint arXiv:2301.04802*.

Alaparthi, S. and Mishra, M., 2020. Bidirectional encoder representations from transformers (bert): A sentiment analysis odyssey. *arXiv preprint arXiv:2007.01127*.

Alshebly, O. Q. and Ahmed, R. M., 2019. Prediction and factors affecting of chronic kidney disease diagnosis using artificial neural networks model and logistic regression model. *Iraqi Journal of Statistical Science*, 28, 1–19.

Alsmadi, M. k., Omar, K. B., Noah, S. A. and Almarashdah, I., 2009. Performance comparison of multi-layer perceptron (back propagation, delta rule and perceptron) algorithms in neural networks. *2009 IEEE International Advance Computing Conference*. URL `http://dx.doi.org/10.1109/iadcc.2009.4809024`.

Andreani, G., Di Fabbrizio, G., Gilbert, M., Gillick, D., Hakkani-Tur, D. and Lemon, O., 2006. Let's discoh: Collecting an annotated open corpus with dialogue acts and reward signals for natural language helpdesks. *2006 IEEE Spoken Language Technology Workshop*, IEEE, 218–221.

Andreassen, S., Riekehr, C., Kristensen, B., Schønheyder, H. C. and Leibovici, L., 1999. Using probabilistic and decision–theoretic methods in treatment and prognosis modeling. *Artificial Intelligence in medicine*, 15 (2), 121–134.

Aziz, H. A., 2017. A review of the role of public health informatics in healthcare. *Journal of Taibah University Medical Sciences*, 12 (1), 78–81.

Bahani, K., Moujabbir, M. and Ramdani, M., 2021. An accurate fuzzy rule-based classification systems for heart disease diagnosis. *Scientific African*, 14, e01019.

Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. *Cornell University - arXiv,Cornell University - arXiv*.

Bandyopadhyay, S., Wolfson, J., Vock, D. M., Vazquez-Benitez, G., Adomavicius, G., Elidrisi, M., Johnson, P. E. and O'Connor, P. J., 2015. Data mining for censored time-to-event data: a bayesian network model for predicting cardiovascular risk from electronic health record data. *Data Mining and Knowledge Discovery*, 29, 1033–1069.

Bayer, M., Kaufhold, M.-A. and Reuter, C., 2022. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55 (7), 1–39.

Beltagy, I., Lo, K. and Cohan, A., 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Beltagy, I., Peters, M. E. and Cohan, A., 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bengio, Y., Ducharme, R. and Vincent, P., 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Bickmore, T. and Giorgino, T., 2006. Health dialog systems for patients and consumers. *Journal of biomedical informatics*, 39 (5), 556–571.

Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T., 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5, 135–146.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al., 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.

Buchanan, B. G. and Shortliffe, E. H., 1984. *Rule based expert systems: the mycin experiments of the stanford heuristic programming project (the Addison-Wesley series in artificial intelligence)*. Addison-Wesley Longman Publishing Co., Inc.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N., 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 1721–1730.

Cavnar, W. B., Trenkle, J. M. et al., 1994. N-gram-based text categorization. *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, Las Vegas, NV, volume 161175, 14.

Chen, T., Xu, R., He, Y. and Wang, X., 2017. Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Systems with Applications*, 72, 221–230.

Chen, Z., Diao, S., Wang, B., Li, G. and Wan, X., 2023. Towards unifying medical vision-and-language pre-training via soft prompts. *arXiv preprint arXiv:2302.08958*.

Chen, Z. and Qian, T., 2018. Transfer capsule network for aspect level sentiment classification. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. URL `http://dx.doi.org/10.18653/v1/p19-1052`.

Cheng, H.-T. and Thoppilan, R., 2022. Lamda: Towards safe, grounded, and high-quality dialog models for everything. *Google AI Blog*.

Cheng, J., Dong, L. and Lapata, M., 2015. Long short-term memory-networks for machine reading. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. URL `http://dx.doi.org/10.18653/v1/d16-1053`.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S. et al., 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Chrimes, D. et al., 2023. Using decision trees as an expert system for clinical decision support for covid-19. *Interactive Journal of Medical Research*, 12 (1), e42540.

Christanti, V. M., Naga, D. S. et al., 2018. Fast and accurate spelling correction using trie and damerau-levenshtein distance bigram. *Telkomnika*, 16 (2), 827–833.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S. et al., 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Conneau, A., Schwenk, H., Barrault, L. and Lecun, Y., 2016. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*.

Cortes, C. and Vapnik, V., 1994. Support-vector networks. *Machine Learning*, 273–297. URL `http://dx.doi.org/10.1023/a:1022627411411`.

Cover, T. and Hart, P., 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13 (1), 21–27.

Dai, H., Liu, Z., Liao, W., Huang, X., Wu, Z., Zhao, L., Liu, W., Liu, N., Li, S., Zhu, D. et al., 2023. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Dai, W., Xue, G.-R., Yang, Q. and Yu, Y., 2007. Transferring naive bayes classifiers for text classification. *AAAI*, volume 7, 540–545.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. and Salakhutdinov, R., 2018. Transformer-xl: Attentive language models beyond a fixed-length context. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. URL `http://dx.doi.org/10.18653/v1/p19-1285`.

Dempster, A. P., Laird, N. M. and Rubin, D. B., 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39 (1), 1–22.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., 2018a. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., 2018b. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North*. URL `http://dx.doi.org/10.18653/v1/n19-1423`.

Dhuliawala, S., Kanojia, D. and Bhattacharyya, P., 2016. Slangnet: A wordnet like resource for english slang. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4329–4332.

Doğan, R. I., Leaman, R. and Lu, Z., 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47, 1–10.

Dziadek, J., Henriksson, A. and Duneld, M., 2017. Improving terminology mapping in clinical text with context-sensitive spelling correction. *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, 235 (241), 8.

Eddy, S. R., 1996. Hidden markov models. *Current opinion in structural biology*, 6 (3), 361–365.

Edwards, C. N., Naik, A., Khot, T., Burke, M. D., Ji, H. and Hope, T., 2023. Synergpt: In-context learning for personalized drug synergy prediction and drug design. *bioRxiv*, 2023–07.

Elfrink, A., Vagliano, I., Abu-Hanna, A. and Calixto, I., 2023. Soft-prompt tuning to predict lung cancer using primary care free-text dutch medical notes. *International Conference on Artificial Intelligence in Medicine*, Springer, 193–198.

Escobar, G. J., Liu, V. X., Schuler, A., Lawson, B., Greene, J. D. and Kipnis, P., 2020. Automated identification of adults at risk for in-hospital clinical deterioration. *New England Journal of Medicine*, 383 (20), 1951–1960.

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T. and Hovy, E., 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Floridi, L. and Chiriatti, M., 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30, 681–694.

Frasconi, P., Soda, G. and Vullo, A., 2002. Hidden markov models for text categorization in multi-page documents. *Journal of Intelligent Information Systems*, 18, 195–217.

Gao, W., Fang, Y., Zhang, F. and Yang, Z., 2020. Representation learning of knowledge graphs using convolutional neural networks. *Neural Network World*, (3).

Genkin, A., Lewis, D. D. and Madigan, D., 2007. Large-scale bayesian logistic regression for text categorization. *technometrics*, 49 (3), 291–304.

Gers, F. A. and Schmidhuber, E., 2001. Lstm recurrent networks learn simple context-free and context-sensitive languages. *IEEE transactions on neural networks*, 12 (6), 1333–1340.

Gers, F. A., Schraudolph, N. N. and Schmidhuber, J., 2002. Learning precise timing with lstm recurrent networks. *Journal of machine learning research*, 3 (Aug), 115–143.

Gillam, S., 2002. Public and private roles in health care systems: reform experience in seven oecd countries. claudia scott.(195 pages, paperback£ 19.99, hardback£ 60.) open university press, 2001. isbn paperback 0-335-20459-7, hardback 0-335-20460-0.

Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., Sakai, S., Seneff, S. and Zue, V., 1995. Multilingual spoken-language understanding in the mit voyager system. *Speech communication*, 17 (1-2), 1–18.

Goldfield, N., Gnani, S. and Majeed, A., 2003. Profiling performance in primary care in the united states. *BMJ*, 326 (7392), 744–747.

Gorry, G. A. and Barnett, G. O., 1968. Experience with a model of sequential diagnosis. *Computers and Biomedical Research*, 1 (5), 490–507.

Gou, B. and Huang, X., 2006. Svm multi-class classification. *Shuju Caiji yu Chuli(Journal of Data Acquisition & Processing)*, 21 (3), 334–339.

Guo, D., Tur, G., Yih, W.-t. and Zweig, G., 2014. Joint semantic utterance classification and slot filling with recursive neural networks. *2014 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 554–559.

Guo, Y., Ge, Y., Yang, Y.-C., Al-Garadi, M. A. and Sarker, A., 2022. Comparison of pretraining models and strategies for health-related social media text classification. *Healthcare*, MDPI, volume 10, 1478.

Gupta, G. and Malhotra, S., 2015. Text document tokenization for word frequency count using rapid miner (taking resume as an example). *Int. J. Comput. Appl*, 975, 8887.

Gupta, V., Lehal, G. S. et al., 2009. A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 1 (1), 60–76.

Hakkani-Tür, D., Tür, G., Celikyilmaz, A., Chen, Y.-N., Gao, J., Deng, L. and Wang, Y.-Y., 2016. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. *Interspeech*, 715–719.

Harris, Z. S., 1954. Distributional structure. *Word*, 10 (2-3), 146–162.

Hayashi, Y., 1990. A neural expert system with automated extraction of fuzzy if-then rules and its application to medical diagnosis. *Advances in neural information processing systems*, 3.

He, K., Zhang, X., Ren, S. and Sun, J., 2015. *Identity Mappings in Deep Residual Networks*. 630–645. URL http://dx.doi.org/10.1007/978-3-319-46493-0_38.

Heckerman, D. E., Horvitz, E. J. and Nathwani, B. N., 1992. Toward normative expert systems: Part i the pathfinder project. *Methods of information in medicine*, 31 (02), 90–105.

Henry, K. E., Hager, D. N., Pronovost, P. J. and Saria, S., 2015. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7 (299), 299ra122–299ra122.

Hinterberger, H., Domingo-Ferrer, J., Kashyap, V., Khatri, V., Snodgrass, R. T., Terenziani, P., Koubarakis, M., Zhang, Y., Joshi, J. B. D., Gamper, J., Böhlen, M. H., Jensen, C. S., Tansel, A. U., Böhlen, M. H., Revesz, P. Z., Mamoulis, N., Wijsen, J., Snodgrass, R. T., Bettini, C., Wang, X. S., Jajodia, S., Dyreson, C. E., Gao, D., Chomicki, J., Toman, D., Shoshani, A., Carlo, C., Snodgrass, R. T., Torp, K., Roddick, J. F., Schiel, U., Silva,

S. F., Grandi, F., Plachouras, V., Lalmas, M., El-Khair, I. A., Carterette, B., Shen, D., Li, H., Ferragina, P., Nitto, I., Zhang, L., Sun, J.-T., Navarro, G., Huang, H., Zhang, B., Moura, E. S. D., Cai, Y., Srinivasan, P., Yan, J., Hu, J., Liu, N., Arenas, M., Breunig, M., Al-Houmaily, Y. J., Samaras, G., Mankovskii, S., George, B., Shekhar, S., Alonso, O., Gertz, M., Montanari, A., Øhrstrøm, P., Hasle, P. F. V., Lorentzos, N. A., Gao, L., Caverlee, J., Jacobsen, H.-A., Marian, A., Hoel, E. G., Felice, P. D., Clementini, E., Kemme, B., Güting, R. H., Vossen, G., Shasha, D., Reuter, A., Alonso, G., Schuldt, H., Moro, M. M., Tsotras, V. J., Manolopoulos, Y., Theodoridis, Y., Fekete, J.-D., Novák, V., Floriani, L. D., Magillo, P., Crochemore, M., Lecroq, T., Despotovic, Z., Agarwal, N., Liu, H., Sion, R., Bonnet, P., Fagin, R., Chen, L., Lechtenbörger, J., Lausen, G. and Amati, G., 2009. Term frequency by inverse document frequency. *Encyclopedia of Database Systems*. URL https://api.semanticscholar.org/CorpusID:38804016.

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural Computation*, 1735–1780. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A. et al., 2022. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35, 30016–30030.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W., 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, Z., Xiang, L., Tu, C., Liu, Z. and Sun, M., 2018. Few-shot charge prediction with discriminative legal attributes. *International Conference on Computational Linguistics,International Conference on Computational Linguistics*.

Huang, K., Altosaar, J. and Ranganath, R., 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Hughes, M., Li, I., Kotoulas, S. and Suzumura, T., 2017. Medical text classification using convolutional neural networks. *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, IOS Press, 246–250.

Irsoy, O. and Cardie, C., 2014. Deep recursive neural networks for compositionality in language.

Jawahar, G., Sagot, B. and Seddah, D., 2018. What does bert learn about the structure of language. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. URL `http://dx.doi.org/10.18653/v1/p19-1356`.

Jia, P., Wang, F. and Xierali, I. M., 2017. Delineating hierarchical hospital service areas in florida. *Geographical Review*, 107 (4), 608–623.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. *European conference on machine learning*, Springer, 137–142.

Joachims, T., 1999. Transductive inference for text classification using support vector machines. *International Conference on Machine Learning,International Conference on Machine Learning*.

Joachims, T., 2001. A statistical learning learning model of text classification for support vector machines. *Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in information retrieval*. URL `http://dx.doi.org/10.1145/383952.383974`.

Johnson, D. E., Oles, F. J., Zhang, T. and Goetz, T., 2001. A decision-tree-based symbolic rule induction system for text categorization. *IBM Systems Journal*, 41 (3), 428–437. URL `http://dx.doi.org/10.1147/sj.413.0428`.

Johnson, R. and Zhang, T., 2015. Semi-supervised convolutional neural networks for text categorization via region embedding. *Cornell University - arXiv,Cornell University - arXiv*.

Johnson, R. and Zhang, T., 2016. Supervised and semi-supervised text categorization using lstm for region embeddings. *Cornell University - arXiv,Cornell University - arXiv*.

Jones, K. S., 1973. Index term weighting. *Information storage and retrieval*, 9 (11), 619–633.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H. and Mikolov, T., 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Kao, H.-C., Tang, K.-F. and Chang, E., 2018. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D., 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kazmierska, J. and Malicki, J., 2008. Application of the naïve bayesian classifier to optimize treatment decisions. *Radiotherapy and Oncology*, 86 (2), 211–216.

Keerthana, T., 2017. Heart disease prediction system using data mining method. *International Journal of Engineering Trends and Technology (IJETT)*, 47 (6), 361–63.

Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C. and Rudzicz, F., 2019. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics*, 100, 100057.

Kim, Y., 2013. Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. URL `http://dx.doi.org/10.3115/v1/d14-1181`.

Kim, Y., 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kim, Y., Lee, H. and Jung, K., 2018. Attnconvnet at semeval-2018 task 1: Attention-based convolutional neural networks for multi-label emotion classification. *arXiv preprint arXiv:1804.00831*.

Kingma, D. P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kitaev, N., Kaiser, Ł. and Levskaya, A., 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Kocbek, S., Cavedon, L., Martinez, D., Bain, C., Mac Manus, C., Haffari, G., Zukerman, I. and Verspoor, K., 2016. Text mining electronic hospital records to automatically classify admissions against disease: measuring the impact of linking data sources. *Journal of Biomedical Informatics*, 64, 158–167.

Kohavi, R. et al., 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. *Kdd*, volume 96, 202–207.

Kononenko, I., 1993. Inductive and bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal*, 7 (4), 317–337.

Kononenko, I., 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23 (1), 89–109.

Koopman, B., Zuccon, G., Nguyen, A., Bergheim, A. and Grayson, N., 2015. Automatic icd-10 classification of cancers from free-text death certificates. *International journal of medical informatics*, 84 (11), 956–965.

Korenius, T., Laurikkala, J., Järvelin, K. and Juhola, M., 2004. Stemming and lemmatization in the clustering of finnish text documents. *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 625–633.

Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., Mascio, A., Zhu, L., Folarin, A. A., Roberts, A. et al., 2021. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117, 102083.

Lafferty, J., McCallum, A. and Pereira, F. C., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lai, S., Xu, L., Liu, K. and Zhao, J., 2022. Recurrent convolutional neural networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*. URL `http://dx.doi.org/10.1609/aaai.v29i1.9513`.

Lamichhane, B., 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R., 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv: Computation and Language,arXiv: Computation and Language*.

Lappenschaar, M., Hommersom, A., Lucas, P. J., Lagro, J. and Visscher, S., 2013. Multilevel bayesian networks for the analysis of hierarchical health care data. *Artificial intelligence in medicine*, 57 (3), 171–183.

Lawrence, R. L. and Wright, A., 2001. Rule-based classification systems using classification and regression tree (cart) analysis. *Photogrammetric engineering and remote sensing*, 67 (10), 1137–1142.

Le, Q. and Mikolov, T., 2014. Distributed representations of sentences and documents. *Cornell University - arXiv,Cornell University - arXiv*.

Ledley, R. S. and Lusted, L. B., 1959. Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science*, 130 (3366), 9–21.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H. and Kang, J., 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36 (4), 1234–1240.

Levin, E., Narayanan, S., Pieraccini, R., Biatov, K., Bocchieri, E., Fabbrizio, G. D., Eckert, W., Lee, S., Pokrovsky, A., Rahim, M. et al., 2000. The at&t-darpa communicator mixed-initiative spoken dialog system. *Sixth International Conference on Spoken Language Processing*.

Lewenberg, Y., Bachrach, Y., Paquet, U. and Rosenschein, J., 2017. Knowing what to ask: A bayesian active learning approach to the surveying problem. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L., 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C. and Lu, Z., 2016a. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Li, T., Shetty, S., Kamath, A., Jaiswal, A., Jiang, X., Ding, Y. and Kim, Y., 2023. Cancergpt: Few-shot drug pair synergy prediction using large pre-trained language models. *arXiv preprint arXiv:2304.10946*.

Li, X., Chen, Y.-N., Li, L., Gao, J. and Celikyilmaz, A., 2017. End-to-end task-completion neural dialogue systems. *arXiv preprint arXiv:1703.01008*.

Li, X., Lipton, Z. C., Dhingra, B., Li, L., Gao, J. and Chen, Y.-N., 2016b. A user simulator for task-completion dialogues. *arXiv preprint arXiv:1612.05688*.

Liang, C., Zhao, Y., Yu, C., Sang, P. and Yang, L., 2022. Hierarchical medical system and local medical performance: A quasi-natural experiment evaluation in shanghai, china. *Frontiers in Public Health*, 10, 904384.

Lieber, O., Sharir, O., Lenz, B. and Shoham, Y., 2021. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*, 1.

Lin, C.-Y. and Hovy, E., 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, 150–157.

Lin, X., He, X., Chen, Q., Tou, H., Wei, Z. and Chen, T., 2019. Enhancing dialogue symptom diagnosis with global attention and symptom graph. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5033–5042.

Lin, Y., Zhao, Z., ZHU, Z., Wang, L., Cheng, K.-T. and Chen, H., 2023. Exploring visual prompts for whole slide image classification with multiple instance learning. *arXiv preprint arXiv:2303.13122*.

Lipsky, A. M. and Lewis, R. J., 2005. Placing the bayesian network approach to patient diagnosis in perspective. *Annals of emergency medicine*, 45 (3), 291–294.

Liu, P., Qiu, X. and Huang, X., 2016. Recurrent neural network for text classification with multi-task learning. *International Joint Conference on Artificial Intelligence,International Joint Conference on Artificial Intelligence*.

Liu, W., Tang, J., Qin, J., Xu, L., Li, Z. and Liang, X., 2020a. Meddg: A large-scale medical consultation dataset for building medical dialogue system.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H. and Wang, P., 2020b. K-bert: Enabling language representation with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2901–2908.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019b. Roberta: A robustly optimized bert pretraining approach. *Cornell University - arXiv,Cornell University - arXiv*.

Lu, H., Ehwerhemuepha, L. and Rakovski, C., 2022. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC medical research methodology*, 22 (1), 181.

Lucas, P., Boot, H. and Taal, B., 1998. Computer-based decision support in the management of primary gastric non-hodgkin lymphoma. *Methods of information in medicine*, 37 (03), 206–219.

Lucas, P. J., De Bruijn, N. C., Schurink, K. and Hoepelman, A., 2000. A probabilistic and decision-theoretic approach to the management of infectious disease at the icu. *Artificial Intelligence in medicine*, 19 (3), 251–279.

Van der Maaten, L. and Hinton, G., 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11).

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S. and Bossan, B., 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Maron, M. E., 1961. Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*, 8 (3), 404–417.

Mawardi, V. C., Susanto, N. and Naga, D. S., 2018. Spelling correction for text documents in bahasa indonesia using finite state automata and levinshtein distance method. *MATEC Web of Conferences*, EDP Sciences, volume 164, 01047.

McCallum, A., Nigam, K. et al., 1998. A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization*, Madison, WI, volume 752, 41–48.

Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C. and Joulin, A., 2017. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J., 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Mousa, A. and Schuller, B., 2017. Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis.

Mujtaba, G., Shuib, L., Idris, N., Hoo, W. L., Raj, R. G., Khowaja, K., Shaikh, K. and Nweke, H. F., 2019. Clinical text classification research trends: systematic literature review and open issues. *Expert systems with applications*, 116, 494–520.

Nan, F., Wang, J. and Saligrama, V., 2015. Feature-budgeted random forest. *International conference on machine learning*, PMLR, 1983–1991.

Nguyen, H. and Nguyen, M.-L., 2017. A deep neural architecture for sentence-level sentiment classification in twitter social networking. *Cornell University - arXiv,Cornell University - arXiv*.

Nii, M., Tsuchida, Y., Kato, Y., Uchinuno, A. and Sakashita, R., 2017. Analysis of classification results for the nursing-care text evaluation using convolutional neural networks. *2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT)*, IEEE, 1–6.

Nori, H., King, N., McKinney, S. M., Carignan, D. and Horvitz, E., 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

O'Donnell, C. A., 2000. Variation in gp referral rates: what can we learn from the literature? *Family practice*, 17 (6), 462–471.

Oleynik, M., Kugic, A., Kasáč, Z. and Kreuzthaler, M., 2019. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. *Journal of the American Medical Informatics Association*, 26 (11), 1247–1254.

OpenAI, 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774. URL `https://api.semanticscholar.org/CorpusID:257532815`.

Ostendorff, M., Bourgonje, P., Berger, M., Moreno-Schneider, J., Rehm, G. and Gipp, B., 2019. Enriching bert with knowledge graph embeddings for document classification. *arXiv preprint arXiv:1909.08402*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. et al., 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.

Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y. and Dehak, N., 2019. Hierarchical transformers for long document classification. *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, IEEE, 838–844.

Patil, R. S., Szolovits, P. and Schwartz, W. B., 1981. Causal understanding of patient illness in medical diagnosis. *Computer-Assisted Medical Decision Making*, Springer, 272–292.

Peng, B., Li, X., Gao, J., Liu, J., Chen, Y.-N. and Wong, K.-F., 2018. Adversarial advantage actor-critic model for task-completion dialogue policy learning. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 6149–6153.

Peng, B., Zhu, C., Li, C., Li, X., Li, J., Zeng, M. and Gao, J., 2020. Few-shot natural language generation for task-oriented dialog. *arXiv preprint arXiv:2002.12328*.

Peng, Y., Yan, S. and Lu, Z., 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Pennington, J., Socher, R. and Manning, C. D., 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L., 2017. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. URL `http://dx.doi.org/10.18653/v1/n18-1202`.

Plisson, J., Lavrac, N., Mladenic, D. et al., 2004. A rule based approach to word lemmatization. *Proceedings of IS*, volume 3, 83–86.

Qin, B., Xia, Y., Prabhakar, S. and Tu, Y., 2009. A rule-based classification algorithm for uncertain data. *2009 IEEE 25th international conference on data engineering*, IEEE, 1633–1640.

Quinlan, J., 1985. Induction of decision trees. *Machine Learning*, 81–106. URL http://dx.doi.org/10.1023/a:1022643204877.

Quinlan, J. R., 1992. C4.5: Programs for machine learning. URL https://api.semanticscholar.org/CorpusID:5262555.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al., 2018. Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al., 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1 (8), 9.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S. et al., 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Ramanayake, R., 2013. Structured printed referral letter (form letter); saves time and improves communication. *Journal of family medicine and primary care*, 2 (2), 145.

Raux, A., 2005. Let's go public! taking a spoken dialog system to the real world. *Proc. of Interspeech 2005*.

Richards, R. J., Hammitt, J. K. and Tsevat, J., 1996. Finding the optimal multiple-test strategy using a method analogous to logistic regression: the diagnosis of hepatolenticular degeneration (wilson's disease). *Medical decision making*, 16 (4), 367–375.

Robertson, S., 2004. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60 (5), 503–520.

Rong, X., 2014. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

Rouillard, A. D., Gundersen, G. W., Fernandez, N. F., Wang, Z., Monteiro, C. D., McDermott, M. G. and Ma'ayan, A., 2016. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, 2016.

Rudnicky, A. I., Thayer, E. H., Constantinides, P. C., Tchou, C., Shern, R., Lenzo, K. A., Xu, W. and Oh, A., 1999. Creating natural dialogs in the carnegie mellon communicator system. *Eurospeech*.

Saif, H., Fernández, M., He, Y. and Alani, H., 2014. On stopwords, filtering and data sparsity for sentiment analysis of twitter. *International Conference on Language Resources and Evaluation*. URL `https://api.semanticscholar.org/CorpusID:1677808`.

Salaba, A. and Chan, L. M., 2023. *Cataloging and classification: an introduction*. Rowman & Littlefield.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T., 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sarikaya, R., Crook, P. A., Marin, A., Jeong, M., Robichaud, J.-P., Celikyilmaz, A., Kim, Y.-B., Rochette, A., Khan, O. Z., Liu, X. et al., 2016. An overview of end-to-end language understanding and dialog management for personal digital assistants. *2016 ieee spoken language technology workshop (slt)*, IEEE, 391–397.

Sarkar, S. and Koehler, J., 2012. A dynamic risk score to identify increased risk for heart failure decompensation. *IEEE Transactions on Biomedical Engineering*, 60 (1), 147–150.

Sarker, A. and Gonzalez, G., 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53, 196–207.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M. et al., 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Schatzmann, J., Thomson, B., Weilhammer, K., Ye, H. and Young, S., 2007. Agenda-based user simulation for bootstrapping a pomdp dialogue system. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, 149–152.

Schneider, K.-M., 2004. A new feature selection score for multinomial naive bayes text classification based on kl-divergence. *Proceedings of the ACL interactive poster and demonstration sessions*, 186–189.

Semigran, H. L., Linder, J. A., Gidengil, C. and Mehrotra, A., 2015. Evaluation of symptom checkers for self diagnosis and triage: audit study. *bmj*, 351.

Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P. and Zue, V., 1998. Galaxy-ii: a reference architecture for conversational system development. *Icslp*, volume 98, 931–934.

Sennrich, R., Haddow, B. and Birch, A., 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Sesen, M. B., Nicholson, A. E., Banares-Alcantara, R., Kadir, T. and Brady, M., 2013. Bayesian networks for clinical decision support in lung cancer care. *PloS one*, 8 (12), e82349.

Shanahan, M., 2022. Talking about large language models. *arXiv preprint arXiv:2212.03551*.

Shen, C., Sun, C., Wang, J., Kang, Y., Li, S., Liu, X., Si, L., Zhang, M. and Zhou, G., 2017. Sentiment classification towards question-answering with hierarchical matching network. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. URL `http://dx.doi.org/10.18653/v1/d18-1401`.

Shi, W., Qian, K., Wang, X. and Yu, Z., 2019. How to build user simulators to train rl-based dialog systems. *arXiv preprint arXiv:1909.01388*.

Shim, H., Hwang, S. J. and Yang, E., 2018. Joint active feature acquisition and classification with variable-size set encoding. *Advances in neural information processing systems*, 31.

Shortliffe, E. H., 1977. Mycin: A knowledge-based computer program applied to infectious diseases. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, American Medical Informatics Association, 66.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A. and Chanona-Hernández, L., 2013. Syntactic dependency-based n-grams as classification features. *Advances in Computational Intelligence: 11th Mexican International Conference on Artificial Intelligence, MICAI 2012, San Luis Potosí, Mexico, October 27–November 4, 2012. Revised Selected Papers, Part II 11*, Springer, 1–11.

Singh, J. and Gupta, V., 2016. Text stemming: Approaches, applications, and challenges. *ACM Computing Surveys (CSUR)*, 49 (3), 1–46.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S. et al., 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Sivarajkumar, S. and Wang, Y., 2022. Healthprompt: A zero-shot learning paradigm for clinical natural language processing. *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, volume 2022, 972.

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V. et al., 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Socher, R., Huval, B., Manning, C. and Ng, A., 2012. Semantic compositionality through recursive matrix-vector spaces. *Empirical Methods in Natural Language Processing,Empirical Methods in Natural Language Processing*.

Socher, R., Pennington, J., Huang, E., Ng, A. and Manning, C., 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. *Empirical Methods in Natural Language Processing,Empirical Methods in Natural Language Processing*.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A. and Potts, C., 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *Empirical Methods in Natural Language Processing,Empirical Methods in Natural Language Processing*.

Song, K., Zeng, X., Zhang, Y., De Jonckheere, J., Yuan, X. and Koehl, L., 2021. An interpretable knowledge-based decision support system and its applications in pregnancy diagnosis. *Knowledge-Based Systems*, 221, 106835.

Soucy, P. and Mineau, G. W., 2001. A simple knn algorithm for text categorization. *Proceedings 2001 IEEE international conference on data mining*, IEEE, 647–648.

Sparck Jones, K., 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28 (1), 11–21.

Suresh, H., Hunt, N., Johnson, A., Celi, L. A., Szolovits, P. and Ghassemi, M., 2017. Clinical intervention prediction and understanding with deep neural networks. *Machine Learning for Healthcare Conference*, PMLR, 322–337.

Sutton, C., McCallum, A. et al., 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4 (4), 267–373.

Tai, K. S., Socher, R. and Manning, C. D., 2014. Improved semantic representations from tree-structured long short-term memory networks. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. URL `http://dx.doi.org/10.3115/v1/p15-1150`.

Tan, S., 2005. Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications*, 28 (4), 667–671.

Tang, K.-F., Kao, H.-C., Chou, C.-N. and Chang, E. Y., 2016. Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning. *NIPS workshop on deep reinforcement learning*.

Tao, Z. and Han, W., 2021. Assessing the impacts of hierarchical healthcare system on the accessibility and spatial equality of healthcare services in shenzhen, china. *ISPRS International Journal of Geo-Information*, 10 (9), 615.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P. and Hashimoto, T. B., 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3 (6), 7.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V. and Stojnic, R., 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. et al., 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tseng, H., Chang, P.-C., Andrew, G., Jurafsky, D. and Manning, C. D., 2005. A conditional random field word segmenter for sighan bakeoff 2005. *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*.

Ubani, S., Polat, S. O. and Nielsen, R., 2023. Zeroshotdataaug: Generating and augmenting training data with chatgpt. *arXiv preprint arXiv:2304.14334*.

Vail, D. L., Veloso, M. M. and Lafferty, J. D., 2007. Conditional random fields for activity recognition. *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, 1–8.

Van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., Kors, J. A. and Furlong, L. I., 2012. The eu-adr corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45 (5), 879–884.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vateekul, P. and Kubat, M., 2009. Fast induction of multiple decision trees in text categorization from large scale, imbalanced, and multi-label data. *2009 IEEE International Conference on Data Mining Workshops*. URL http://dx.doi.org/10.1109/icdmw.2009.94.

Velikova, M., Van Scheltinga, J. T., Lucas, P. J. and Spaanderman, M., 2014. Exploiting causal functional relationships in bayesian network modelling for personalised healthcare. *International Journal of Approximate Reasoning*, 55 (1), 59–73.

Venkataraman, G. R., Pineda, A. L., Bear Don't Walk IV, O. J., Zehnder, A. M., Ayyar, S., Page, R. L., Bustamante, C. D. and Rivas, M. A., 2020. Fastag: Automatic text classification of unstructured medical narratives. *PLoS one*, 15 (6), e0234647.

Verduijn, M., Peek, N., Rosseel, P. M., de Jonge, E. and de Mol, B. A., 2007. Prognostic bayesian networks: I: Rationale, learning procedure, and clinical use. *Journal of Biomedical Informatics*, 40 (6), 609–618.

Verma, T., Renu, R. and Gaur, D., 2014. Tokenization and filtering process in rapidminer. *International Journal of Applied Information Systems*, 7 (2), 16–18.

Vila-Francés, J., Sanchis, J., Soria-Olivas, E., Serrano, A. J., Martinez-Sober, M., Bonanad, C. and Ventura, S., 2013. Expert system for predicting unstable angina based on bayesian networks. *Expert systems with applications*, 40 (12), 5004–5010.

Walker, M. A., Aberdeen, J. S., Boland, J. E., Bratt, E. O., Garofolo, J. S., Hirschman, L., Le, A. N., Lee, S., Narayanan, S. S., Papineni, K. et al., 2001. Darpa communicator dialog travel planning systems: the june 2000 data collection. *INTERSPEECH*, 1371–1374.

Walker, M. A., Rudnicky, A. I., Aberdeen, J. S., Bratt, E. O., Garofolo, J. S., Hastie, H. W., Le, A. N., Pellom, B. L., Potamianos, A., Passonneau, R. J. et al., 2002. Darpa communicator evaluation: progress from 2000 to 2001. *Interspeech*, 273–276.

Wang, Y., Deng, J., Wang, T., Zheng, B., Hu, S., Liu, X. and Meng, H., 2023a. Exploiting prompt learning with pre-trained language models for alzheimer's disease detection. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 1–5.

Wang, Y., Huang, M., zhu, x. and Zhao, L., 2015. Attention-based lstm for aspect-level sentiment classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. URL `http://dx.doi.org/10.18653/v1/d16-1058`.

Wang, Y., Sun, A., Han, J., Liu, Y. and Zhu, X., 2017a. Sentiment analysis by capsules. *Proceedings of the 2018 World Wide Web Conference on World*

Wide Web - WWW '18. URL http://dx.doi.org/10.1145/3178876.3186015.

Wang, Y., Wang, Y., Peng, Z., Zhang, F., Zhou, L. and Yang, F., 2023b. Medical text classification based on the discriminative pre-training model and prompt-tuning. *Digital Health*, 9, 20552076231193213.

Wang, Y.-Y., Deng, L. and Acero, A., 2011. Semantic frame-based spoken language understanding. *Spoken language understanding: systems for extracting semantic information from speech*, 41–91.

Wang, Z., Hamza, W. and Florian, R., 2017b. Bilateral multi-perspective matching for natural language sentences. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. URL http://dx.doi.org/10.24963/ijcai.2017/579.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D. et al., 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.

Wei, J. and Zou, K., 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Wei, Z., Liu, Q., Peng, B., Tou, H., Chen, T., Huang, X.-J., Wong, K.-F. and Dai, X., 2018. Task-oriented dialogue system for automatic diagnosis. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 201–207.

Wen, T.-H., Gasic, M., Mrksic, N., Su, P.-H., Vandyke, D. and Young, S., 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Wen, T.-H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S. and Young, S., 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Whitney, D. L. and Evans, B. W., 2010. Abbreviations for names of rock-forming minerals. *American mineralogist*, 95 (1), 185–187.

Wiens, J., Guttag, J. and Horvitz, E., 2016. Patient risk stratification with time-varying parameters: a multitask learning approach. *The Journal of Machine Learning Research*, 17 (1), 2797–2819.

Wu, J., Xie, R., Liu, Z. and Sun, M., 2016. Knowledge representation via joint learning of sequential text and knowledge graphs. *arXiv preprint arXiv:1609.07075*.

Wu, X., Lv, S., Zang, L., Han, J. and Hu, S., 2019. Conditional bert contextual augmentation. *Computational Science–ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part IV 19*, Springer, 84–95.

Xie, R., Liu, Z., Jia, J., Luan, H. and Sun, M., 2016. Representation learning of knowledge graphs with entity descriptions. *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Xu, J., Chen, K., Qiu, X. and Huang, X., 2016. Knowledge graph representation with jointly structural and textual encoding. *arXiv preprint arXiv:1611.08661*.

Xu, L., Zhou, Q., Gong, K., Liang, X., Tang, J. and Lin, L., 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 7346–7353.

Xu, P. and Sarikaya, R., 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling. *2013 ieee workshop on automatic speech recognition and understanding*, IEEE, 78–83.

Xu, Z., Kusner, M., Weinberger, K. and Chen, M., 2013. Cost-sensitive tree of classifiers. *International conference on machine learning*, PMLR, 133–141.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. and Le, Q., 2019a. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv: Computation and Language,arXiv: Computation and Language*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. and Le, Q. V., 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E., 2015. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. URL `http://dx.doi.org/10.18653/v1/n16-1174`.

Yao, K., Peng, B., Zhang, Y., Yu, D., Zweig, G. and Shi, Y., 2014. Spoken language understanding using long short-term memory neural networks. *2014 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 189–194.

Yao, K., Zweig, G., Hwang, M.-Y., Shi, Y. and Yu, D., 2013. Recurrent neural networks for language understanding. *Interspeech*, 2524–2528.

Yet, B., Bastani, K., Raharjo, H., Lifvergren, S., Marsh, W. and Bergman, B., 2013. Decision support system for warfarin therapy management using bayesian networks. *Decision Support Systems*, 55 (2), 488–498.

Yi, K. and Beheshti, J., 2009. A hidden markov model-based text classification of medical documents. *Journal of Information Science*, 35 (1), 67–81.

Young, S., 2006. Using pomdps for dialog management. *2006 IEEE Spoken Language Technology Workshop*, IEEE, 8–13.

Young, S., Gašić, M., Thomson, B. and Williams, J. D., 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101 (5), 1160–1179.

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X. et al., 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Zhang, M.-L. and Zhang, K., 2010. Multi-label learning by exploiting label dependency. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 999–1008.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V. et al., 2022a. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, W., Yoshida, T. and Tang, X., 2008. Tfidf, lsi and multi-word in information retrieval and text categorization. *2008 IEEE International Conference on Systems, Man and Cybernetics*, IEEE, 108–113.

Zhang, X., Zhao, J. and LeCun, Y., 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zhang, Y. and Chen, D. Z., 2023. Gpt4mia: Utilizing geneative pre-trained transformer (gpt-3) as a plug-and-play transductive model for medical image analysis. *arXiv preprint arXiv:2302.08722*.

Zhang, Y., Jiang, Z., Zhang, T., Liu, S., Cao, J., Liu, K., Liu, S. and Zhao, J., 2020a. Mie: A medical information extractor towards medical dialogues. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6460–6469.

Zhang, Y., Zhang, Y., Qi, P., Manning, C. D. and Langlotz, C. P., 2021. Biomedical and clinical english model packages for the stanza python nlp library. *Journal of the American Medical Informatics Association*, 28 (9), 1892–1899.

Zhang, Z., Takanobu, R., Zhu, Q., Huang, M. and Zhu, X., 2020b. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63 (10), 2011–2027.

Zhang, Z., Zhang, A., Li, M. and Smola, A., 2022b. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Zhou, M., Li, Z., Tan, B., Zeng, G., Yang, W., He, X., Ju, Z., Chakravorty, S., Chen, S., Yang, X. et al., 2021. On the generation of medical dialogs for covid-19. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Zhou, X., Wan, X. and Xiao, J., 2015. Attention-based lstm network for cross-lingual sentiment classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. URL `http://dx.doi.org/10.18653/v1/d16-1024`.

Zhu, W., Zhou, R., Yuan, Y., Timothy, C., Jain, R. and Luo, J., 2023. Segprompt: Using segmentation map as a better prompt to finetune deep models for kidney stone classification. *arXiv preprint arXiv:2303.08303*.

Zhu, X., Sobihani, P. and Guo, H., 2015. Long short-term memory over recursive structures. *International Conference on Machine Learning,International Conference on Machine Learning*.