2021

Computational Intelligence for the Micro Learning

Jiayin Lin

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Computational Intelligence for the Micro Learning

Jiayin Lin

Supervisors:
Associate Professor Jun Shen
Dr Tingru Cui
Associate Professor Ping Yu

This thesis is presented as part of the requirement for the conferral of the degree:
Doctor of Philosophy

This research has been funded by the support of scholarship through Discovery Project grant

(DP180101051) of the Australian Research Council

University of Wollongong

School of Computing and Information Technology

July 2021

Abstract

The developments of the Web technology and the mobile devices have blurred the time and space boundaries of people's daily activities, which enable people to work, entertain, and learn through the mobile device at almost anytime and anywhere. Together with the life-long learning requirement, such technology developments give birth to a new learning style, **micro learning**. Micro learning aims to effectively utilise learners' fragmented spare time and carry out personalised learning activities.

However, the massive volume of users and the online learning resources force the micro learning system deployed in the context of enormous and ubiquitous data. Hence, manually managing the online resources or user information by traditional methods are no longer feasible. How to utilise computational intelligence based solutions to automatically managing and process different types of massive information is the biggest research challenge for realising the micro learning service. As a result, to facilitate the micro learning service in the big data era efficiently, we need an intelligent system to manage the online learning resources and carry out different analysis tasks. To this end, an intelligent micro learning system is designed in this thesis.

The design of this system is based on the service logic of the micro learning service. The micro learning system consists of three intelligent modules: learning material pre-processing module, learning resource delivery module and the intelligent assistant module. The pre-processing module interprets the content of the raw online learning resources and extracts key information from each resource. The pre-processing step makes the online resources ready to be used by other intelligent components of the system. The learning resources delivery module aims to recommend personalised learning resources to the target user base on his/her implicit and explicit user profiles. The goal of the intelligent assistant module is to provide some evaluation or assessment services (such as student dropout rate prediction and final grade prediction) to the educational resource providers or instructors. The educational resource providers can further refine or modify the learning materials based on these assessment results.

Before diving into the technical details of the proposed models, in this thesis, the status of current research datasets is analysed and discussed. Based on the prior works from the related area, we identify several data challenges that impede the development of online learning research. The requirements of the research dataset for this area are highlighted.

With the big data challenges mentioned before, in order to fit the big data context, all the proposed solutions are intelligence based models constructed by different task-specific datasets. Before demonstrating the proposed solutions, recent studies about the state-of-the-art models for different tasks are reviewed and discussed.

For pre-processing the raw learning resources, an information extraction model and a content analysis model are proposed in this research. The proposed content analysis model is designed to interpret the short informal text of the online learning resource, such as the users' comments and learners' discussions. The proposed information extraction model aims to extract key information (such as keywords, name entities, and relationships) from a given sequence of inputs. The extracted information is fine-grained, which can be directly used by other components of the system.

For the task of learning resource recommendation, a deep-cross-attention network is designed to boost recommendation performance. Through integrating various types of the neural networks, the proposed model can effectively combine three functionalities: generating high-order feature interactions, distinguishing the importance differences of different features, and maintaining the original input information for the deep neural network. In the intelligent assistant module, we design a student dropout rate prediction model to detect whether a student is about to drop out from an enrolled online course. The proposed model is in the double-tower structure, which can separately process different types and granularities of information.

All the models mentioned above are constructed by different task-specific real-world datasets collected from various application scenarios in different chapters. Also, the performance of each model is compared with the state-of-the-art solutions through comprehensive experiments. The experiment results show that most of the proposed models in this research outperform the baselines.

Moreover, to further enhance the generalisation capability of the recommender system, a GAN-based optimisation strategy is also proposed in this research. With the pilot experiment, the proposed new optimisation strategy is verified to be useful and efficient for the micro learning service. We then further propose a novel GAN framework (which contains three generators and one discriminator) for this optimisation strategy. We also carefully design several loss functions to further refine the optimisation procedure. Through a comprehensive experiment, the proposed GAN framework and the loss functions are proved to be effectiveness.

Acknowledgments

Every time when I look back, I am always surprised on the long and often tedious research journey that I have already walked through. Many plans I made for this journey did not go well, and maybe I will not become the person I planned to be, but this journey itself is very memorable. Not trying to be humble, the accomplishment in research might not make this journey memorable, but the people around me make this journey memorable. In the near future, or maybe when I leave academia, I will definitely forget the exact titles of the papers I have published during my PhD, but I do not think I will forget the people who helped me and accompanied me during this period. Hence, first and foremost, at the very beginning of this thesis, let me express my sincere appreciation to these people.

Firstly, I would like to thank my parents, my dad Jianghuai Lin and my mom Zhuang Lin. I am very grateful they could understand my choice to start this journey, even though they had no idea how to do research and what I was about to do research. The only thing they knew was it would not be an easy job. They supported me unwaveringly all the time. And this support was my most powerful backing all the way along this journey. Without this understanding and support, I would not even start my academia career. Also, I want to thank my aunt, Mrs Lin Jiang, my uncle, Dr Hui Lu, and my cousin, Dr Sijie Lu, thanks for their good care. They gave me a place where I could call it home when I was far away from my homeland.

Then, I would like to thank my principal supervisor A/Prof. Jun Shen and two co-supervisors, Dr Tingru Cui and A/Prof. Ping Yu. It is them who gave me this chance to start my research. They provided me with guidance letting me know how to become a qualified researcher. Also, thanks to Dr Geng Sun, he helped me a lot with my research all the time. I also gratefully acknowledge the Australian Government Research scholarship for giving me financial support, which made my research possible.

This journey is not merely about the academic; the mundane daily life is also one part of it. Hence, I also want to express my gratefulness to my friends. Dr Ting Song offered me countless helps when I lived in Wollongong. Even when I got stuck in China during the COVID-19 epidemic, he helped me to deal with the trivia I left in Australia. Mr Zhexuan Zhou, another close friend of mine, we share similar research interests. In addition to discussing interesting AI topics, we went to the gym together,

played basketball together, and hung out at the weekend together. My after-school life would definitely be much dimmer without him. Miss Yunshu Zhu, who helped me a lot and brought much laugh to us when we were working in the lab. Dr Jianqing Wu, a warm-hearted peer, showed me around the campus and the Wollongong city when I first arrived. Without his help, I could not get used to the new environment so quickly.

Many thanks to those old friends I made in UESTC, UniMelb, and Alibaba Inc. Even though we seldom met these years, but they always offered their support in the distance. Some of them gave me technical support, and some of them were patient listeners who were never tired of my boring research ideas and tedious stories about university life. Sorry, due to the space limit, I cannot list your names one by one with unforgettable stories. So, thank you all.

Thanks to those scholars from different journals and conferences review system, who have ever rejected my submissions. These rejections made me feel frustrated and maybe angry sometimes, but your feedback and comments let me see my shortcomings and urged me to work harder and move forward.

Thanks to those ones who are or will read my thesis, thank you for being the last audience for my PhD research journey.

Lastly, I would like to express my special thanks to my powerful country, China.

At this moment, I think it is time to say goodbye to the old days. It is not the end but the beginning of the next step. I have no idea after I graduate, where I will be, what kind of job I will get, and whether I will still be active in academia. Nothing is settled yet! If life is not like a box of chocolate, what's the fun? There is a saying that "*today is difficult, tomorrow will be difficult, but the day after tomorrow will be beautiful*". I believe the difficulties I meet are the dots, and some day there will be a line connecting these dots. By that time, when I look back, everything will look reasonable and worthwhile. Many people call this destiny; I call it life.

Certification

I, Jiayin Lin, declare that this thesis submitted in fulfilment of the requirements for the conferral of the degree Doctor of Philosophy, from the University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

_____

<<*Jiayin Lin*>>

<<Date (26th July 2021)>>

List of Names and Abbreviations

Name	Abbreviation
Technology-enhanced Learning	TEL
Artificial Intelligence	A.I.
Multi-layer Perceptron	MLP
Long Short-Term Memory	LSTM
Gated Recurrent Unit	GRU
Generative Adversarial Network	GAN
Rectified Linear Unit	ReLU
Massive Open Online Course	MOOC
Open Educational Resource	OER
Normalized Discounted Cumulative Gain	nDCG
Mean Reciprocal Rank	MRR
Area Under Curve	AUC
Computer Vision	CV
Natural Language Processing	NLP
Convolutional Neural Network	CNN
Recurrent Neural Network	RNN
Mean Square Error	MSE
Conditional Random Field	CRF
Bidirectional Encoder Representations from Transformers	Bert
Optical Character Recognition	OCR
Automatic Speech Recognition	ASR
Ant Colony Optimisation	ACO
Particle Swarm Optimization	PSO

Genetic Algorithm	GA
Reinforcement Learning	RL

Table of Contents

List of Figures

List of Tables

Publications

Published Papers:

1. **Jiayin Lin**, Geng Sun, Jun Shen, David Pritchard, Tingru Cui, Dongming Xu, Li Li, Ghassan Beydoun, and Shiping Chen: 'Deep-Cross-Attention Recommendation Model for Knowledge Sharing Micro Learning Service'. In. International Conference on Artificial Intelligence in Education pp. 168-173, 2020. (CORE: A)
2. **Jiayin Lin**, Geng Sun, Jun Shen, Tingru Cui, David Pritchard, Dongming Xu, Li Li, Wei Wei, Ghassan Beydoun, and Shiping Chen: 'Attention-Based High-Order Feature Interactions to Enhance the Recommender System for Web-Based Knowledge-Sharing Service'. In. International Conference on Web Information Systems Engineering pp. 461-473, 2020. (CORE: B)
3. **Jiayin Lin**: 'Hybrid Translation and Language Model for Micro Learning Material Recommendation'. In. 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT) pp. 384-386, 2020. (CORE: B)
4. **Jiayin Lin**, Zhexuan Zhou, Geng Sun, Jun Shen, David Pritchard, Tingru Cui, Dongming Xu, Li Li, and Ghassan Beydoun: 'Deep Sequence Labelling Model for Information Extraction in Micro Learning Service'. In. 2020 International Joint Conference on Neural Networks (IJCNN) pp. 1-10, 2020. (CORE: B)
5. **Jiayin Lin**, Geng Sun, Tingru Cui, Jun Shen, Dongming Xu, Ghassan Beydoun, Ping Yu, David Pritchard, Li Li, and Shiping Chen: 'From ideal to reality: segmentation, annotation, and recommendation, the vital trajectory of intelligent micro learning', World Wide Web, 2019, 23, (3), pp. 1-21. (CORE: A)
6. **Jiayin Lin**, Geng Sun, Jun Shen, Tingru Cui, Ping Yu, Dongming Xu, and Li Li: 'A Survey of Segmentation, Annotation, and Recommendation Techniques in Micro Learning for Next Generation of OER'. In. 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD) pp. 152-157, 2019. (CORE: B)
7. **Jiayin Lin**, Geng Sun, Jun Shen, Tingru Cui, Ping Yu, Dongming Xu, Li Li, and Ghassan Beydoun: 'Towards the readiness of learning analytics data for micro learning'. In. International Conference on Services Computing pp. 66-76, 2019. (CORE: B)

Papers Under Review:

8. **Jiayin Lin**, Geng Sun, Jun Shen, Tingru Cui, David Pritchard, Ping Yu, Ghassan Beydoun, Li Li, Applying GAN Techniques in Recommenders for Micro Learning, *IEEE Transactions on Learning Technologies*
9. **Jiayin Lin**, Geng Sun, Jun Shen, Ghassan Beydoun, David Pritchard, Ping Yu, Tingru Cui, Li Li, Dongming Xu, False Negative, False Positive, and Non-positive Feedback: Adversarial Learning for Recommendation in Knowledge Sharing Services, *Neurocomputing*
10. **Jiayin Lin**, Geng Sun, Jun Shen, David Pritchard, Ping Yu, Tingru Cui, Li Li, Ghassan Beydoun, MOOC Student Dropout Rate Prediction via Separating and Conquering Micro and Macro Information, *International Conference on Neural Information Processing, 2021*
11. **Jiayin Lin**, Geng Sun, Jun Shen, David Pritchard, Ping Yu, Tingru Cui, Dongming Xu, Li Li, Ghassan Beydoun, From Computer Vision to Short Text Understanding: Applying Similar Approaches into Different Disciplines, *International Conference on Neural Information Processing, 2021*

This thesis is in the form of thesis by complication. The content of this thesis is based on the above publication with a slight modification and adjustment. I, Jiayin Lin, am the first author of the above publication and I am the major contributor of the above work. For each work, my contributions include all aspects of research design, experimenting, and paper composing.

For each chapter of this thesis:

Chapter 2 is mainly based on the publication 5, and 6, and partially based on the related work section of publication 1,2,3,4,8,9,10,11.

Chapter 3 is mainly based on the publication 7.

Chapter 4 is mainly based on the publication 4 and 11.

Chapter 5 is mainly based on the publication 1, and 2 and partially based on the publication 3.

Chapter 6 is mainly based on the publication 10.

Chapter 7 is mainly based on the publication 8 and 9.

Other co-authored publications which are not included in this thesis:

1. Geng Sun, **Jiayin Lin**, Jun Shen, Tingru Cui, Dongming Xu, and Huaming Chen: 'Evolutionary Learner Profile Optimization Using Rare and Negative Association Rules for Micro Open Learning'. In. International Conference on Intelligent Tutoring Systems pp. 432-

440, 2020. (CORE: B)

2. Geng Sun, **Jiayin Lin**, Jun Shen, Tingru Cui, Dongming Xu, and Mahesh Kayastha: ‘Refinement and augmentation for data in micro open learning activities with an evolutionary rule generator’, British Journal of Educational Technology, 2020, 51, (5), pp. 1843-1863

Chapter 1 Introduction

1.1 Background

The soaring development of the Internet and the mobile device catalyse the evolution of various mobile applications and services. Such a development trend aims to break the time-space restriction, such that people can work, entertain, or study at almost any time and anywhere. Moreover, in the meanwhile, due to the fast-paced modern life and the immersed usage of mobile devices, people's spare time is split into small irregular time-fragments between the switch of different activities. After the entertainment industry first starts utilising the high value of people's fragmented time, researchers in the area of technology-enhanced learning (TEL) start to investigate how to make good use of such small chunks of time and carry out effective and personalised learning activities.

As discussed in one prior study [Sun, et al., 2018], to fit the user's fast-paced lifestyle and satisfy the lifelong learning requirement, it is necessary to deliver users personalised adaptive small chunks of learning materials. These small chunks of learning materials are coherent knowledge points and can be learnt in relatively short and isolated time duration. The term 'micro learning' used in this thesis refers to the novel learning style that mainly utilises small-time fragments with small pieces of learning materials. As found in the studies [Guo, et al., 2014, Anderson, et al., 2014], users' engagement in online learning activities plunges quickly after 7 minutes, and videos with a short time duration are more popular among learners. Moreover, as pointed out in the study [Syeda-Mahmood, et al., 2001], for a short learning session such as a short video, users are less likely to leave out the knowledge points. Hence, micro learning has great potential in fitting the fast-paced lifestyle and alleviating engagement or dropout issues.

Based on the service logic of micro learning [Lin, et al., 2019], the full framework can be divided into three modules: learning material pre-processing, learning material delivery and the intelligent assistant modules. The first module aims to pre-process the online learning materials and make them ready to be further used by other intelligent models of the system. The second module is mainly focused on recommending learner with personalised materials based on his/her user-characteristics, such as historical learning activities, knowledge level, and learning preference. The last module is for providing some intelligent services or functionalities (like dropout rate prediction for early

intervention of low engagement) for the educational resource providers (such as instructors or institutions). For a real-time personalised learning service in the context of big data, the above mentioned three modules should be highly automatic, modularised, and cohesive.

1.2 Research Problems

In the big data era, the explosive growth of the volume of learning resource on the Internet floods learners with tremendous information in both variety and complexity. In the meantime, after leading universities proposed the idea of open educational resource (OER), online learning is promoted by more and more institutions and companies as a new learning style. Also, many existing online learning platforms such as edX[1] and Coursera[2] are designed to be used under the context of big data with massive digital learning resource and users. Managing massive online learning resource and users' information manually by traditional methods is significantly labour intensive and costly. The workload of manually processing online information is beyond imagination. Therefore, intelligence, automaticity, and personalisation become the three key factors to a successful e-learning application. How to use advanced artificial intelligence techniques to maintain the micro learning system and manage learning resource and online users is critical to the development of online learning. Hence, now and in the future, for the service of online learning, it should be deployed on or attached to a highly intelligent system that could automatically manage the massive learning resource and analysis underlying characteristics of online learners and resources.

Specifically, for an open micro learning platform, massive learning resources are generated daily, and online learners come and go every second. Manually pre-process the new generated micro learning resource one by one is deemed to be impossible under the context of big data. Moreover, the massive users in a learning platform have quite different knowledge backgrounds, learning requirements, and learning preferences. Hence, the analysis of users' information and recommending suitable learning resources to the target users should also be automatic, it is unrealistic to recruit a large number of educationalists to do this job manually. Lastly, the educational providers are eager to have an insight view of the effectiveness of the provided online resources, which requires the micro learning system has relevant intelligent modules to analyse the learning activities and give those feedbacks. Generally speaking, it strongly demands a system that can automatically and intelligently

[1] https://www.edx.org/

[2] https://www.coursera.org/

the tasks mentioned above for the micro learning service. Artificial intelligence (A.I.) enhanced micro learning is certain to be a future research trend.

Based on the three modules of the micro learning system, my research aims to solve the problem of the lack of comprehensive design of intelligent modules for the micro learning service system. In my research, I mainly focus on filling the following three gaps:

1. The lack of pre-processing strategies for micro learning resource.
2. The lack of an effective recommender system for the proposed micro learning system.
3. The lack of assistant plugins can provide the educational resources providers with an insight view of the effectiveness of the provided online learning resources.

1.3 Research Objectives and Contributions

From a high perspective, the main research object is to design a micro learning system framework and develop some task-specific A.I. techniques.

As the micro learning service involves a series of complex data processing and analysing stages, the complete micro learning service could be divided into several modules and each of which has a different obligation. The effectiveness of the micro learning service is determined by each of the processing or analysing module. Due to the different data requirements and different parts of a micro learning system, we should apply different A.I. techniques for different components. Based on the resource we have and my research background, besides the main research objective of system design, there are three other sub-objectives:

1. Designing an intelligent pre-processing module that aims to make the learning material ready to be delivered to the target learners. The pre-processing step should be able to extract different perspectives of hidden information from the learning resource. This module is made up of different A.I. models such as information extraction and text analysis.
2. Designing a recommender system for delivering online learning resource to the target learners. For personalised recommendation of micro learning resource: the recommendation strategy should be based on analysing users' historical learning records, finding similar learners and mining user’s learning requirements and knowledge level.
3. Designing a dropout rate prediction model for assisting educational resource providers in evaluating the effectiveness of learning activities. The dropout rate prediction model can be

further used for the early intervention of students with high dropout intention.

As the results of this research, the contribution of this research can be highlighted as bellows.

1. A micro learning system framework that focuses on depicting how the A.I. techniques can be used to boost the micro learning service. This framework also demonstrates the workflow of the involved intelligent modules and the relationships among these modules (in Chapter 3).
2. The technical details of several intelligent models I designed during this research project. These models involve different business aspects (such as backend information pre-processing, service evaluation, and resource delivery) of the micro learning system. These proposed models have significant reference value to the future research of online learning and micro learning (in Chapters 4,5,6, and 7).
3. A comprehensive discussion and analysis of current mainstream public datasets for the online learning research is provided (in Chapter 2 and 3).

1.4 Outline of the Thesis

The rest of the thesis is organised as follows:

Chapter 2 is introducing the basic concepts of this research together with the literature review of the current status of the related research. The definition, micro learning characteristics, and its relationships with traditional online learning are first introduced and discussed. Then the recent studies of content understanding (such as information extraction and text analysis) are presented. This part is related to the pre-processing stage of the micro learning service. After that, recent studies of the recommender system are discussed. This part contains the current recommendation strategies for online learning, the mainstream recommender systems, and the differences between micro learning recommendation and the recommenders in other domains. Lastly, related work about student dropout rate prediction is introduced in this chapter.

Chapter 3 is presenting the big picture of the technical part of this research. The A.I. enhanced framework of the proposed micro leaning system is presented and introduced in this chapter. Each component of the framework is explained in detail, including the functionalities and relationships. Also, the data source of each intelligent component is introduced in this chapter, including the characteristics of each type of data and its workflow.

The detail of how the micro learning resources should be pre-processed is demonstrated in chapter 4.

This chapter consists of two parts, one is about information extraction, and another one is about text analysis. For each part, the technical aspects of the proposed intelligent model are introduced. In the experimental section of this chapter, the configurations of the proposed models are demonstrated, and the proposed models are compared with the state-of-the-art solutions through comprehensive experiments. Moreover, the statistical information of the datasets used in the experiments is also discussed and analysed in this chapter.

Chapter 5 is focusing on the recommendation strategy of the micro learning service. First, the proposed recommendation strategy is introduced and discussed in this chapter in detail. Then, in order to demonstrate the effectiveness of the proposed recommendation strategy, the proposed solution is compared with the state-of-the-art recommender models through experiments. At the end of this chapter, together with the involved dataset, the experiment results are analysed and discussed.

Chapter 6 is investigating the online students' dropout rate. The dropout rate prediction model is one of the representative intelligent assistant plugins for the proposed micro learning system. Similar to other A.I. based model of this system, the technical detail of the proposed solution is introduced first. Then we conduct a comprehensive experiment in which the proposed model is compared with the state-of-the-art solutions. The experiment results are analysed and summarised at the end of this chapter.

Chapter 7 is a further investigation of the optimisation strategy for the recommender system used in the micro learning system. Firstly, the current gaps of micro learning resource recommendation are discussed and analysed. The proposed generative adversarial network-based optimisation strategy for the micro learning recommender system is introduced. Then, the experiment of comparing the proposed optimisation strategy with the conventional optimisation method is demonstrated in this chapter. Lastly, the experiment results are comprehensively analysed.

This thesis is concluded in Chapter 8. The suggestion of future research about micro learning is also given in this chapter.

Chapter 2 Literature Review

In this chapter, we discuss the related studies of the micro learning service. For each intelligent module of the micro learning system, we review and discuss the works of information extraction, text analysis (NLP techniques), recommender system, and student dropout rate prediction. For the GAN-based optimisation strategy of the recommender system, we also investigate the related work of the GAN technique and the GAN-based recommendation framework.

2.1 Micro Learning

2.1.1 The Light-Weight Learning Service

As mentioned in the previous chapter, to fit the user's fast-paced lifestyle and satisfying the life-long learning requirements, it is necessary to make good use of the learners' small chunks of spare time to carry effective learning activities. This learning requirement gives birth to a novel learning style, micro learning. This learning style aims to utilise learner's fragmented spare time and curry out effective learning activities. The term 'micro' refers to the short time length that a learner needs to take to consume a learning material. A micro learning material is also in a small format which could be just a single knowledge point or a subsection of a learning topic. The term 'learning' refers to the online learning service or activity that a user is involved in. The online learning service or learning activity could be formal, informal, or non-formal.

Moreover, the soaring development of the Internet and mobile device catalyse the evolution of the mobile application and service. Such development breaks the time-space restriction and makes it possible that people can work, entertain, and study almost anytime and anywhere. Highlighted in one early study that the development of the Web and mobile platforms are significant factors for enabling novel learning methods such as micro learning [Kovachev, et al., 2011]. This research also indicated that the micro learning consists of a fast, convenient and instant capture of the self-identified knowledge gaps, understanding them with the help of online resources, creation of a learning object out of these online resources and integration of that learning object into small learning activities interwoven into our daily life [Kovachev, et al., 2011]. But the researchers of this study only focused

on promoting micro learning as an informal learning method. As discussed in [Bruck, et al., 2012] , micro learning has been designed to increase the usage of technology in learning by making it more convenient and adapting the e-learning to the fact that users often have considerable difficulty make time for long stretches of learning activity outside the dedicated study times and institutional programs of schools, colleges or universities.

Overall, the micro learning service has great potential in fitting the fast-paced lifestyle and alleviating engagement issues. As concluded in one prior study [Mohammed, et al., 2018], that using micro learning techniques, the effectiveness and efficiency of learning could be improved, and the knowledge could stay memorable for a longer period.

2.1.2 Challenges in the Micro Learning Service

2.1.2.1 Content Analytics in Micro Learning

As discussed before, a great portion of online learning resources are in the audio and video format. However, there is a huge research gap in interpreting and analysing the video or audio content, especially for the micro learning materials. Nowadays most of the studies heavily rely on OCR, ASR, NLP techniques, which are indirect approaches to interpreting and analysing the content of the learning resources. The combination of OCR, ASR, and NLP aims to transfer all audio and visual information into the textual form. Due to the technical difficulties in directly interpreting the content of the video stream or audio signal, textual information becomes the only remaining metadata that researchers can interpret and analyse. As discussed in [Bolettieri, et al., 2007], although automatically-generated metadata and annotations by using ASR or OCR are sometimes error-prone, they are practically two of the limited options for researchers to make the audio-visual content retrievable and accessible.

2.1.2.2 Trade-off between Personalisation and System Workload

When a real-life problem involves user interactions, machine learning techniques are frequently used to construct the user modelling to represent the user's profile and other related information that matters in different complex scenarios. Such a model is intended to represent general user information. Conventionally, for the instances of a user model, all users could have different feature values but share the same set of feature weightings. For a personalised online learning service, how nicely the user information can be modelled and represented determines the extent of personalisation

service can a system offer. Many studies argue that, to boost the personalisation of an recommender system in the educational scenario, it is necessary to construct a user model for each user individually [Al-Shamri, et al., 2008, Wasid, et al., 2015, Fong, et al., 2008, Bobadilla, et al., 2011, Ujjin, et al., 2003, Ujjin, et al., 2002]. As discussed in [Al-Shamri, et al., 2008], the main features reflecting different users' preference are naturally different; for example, some users mainly rely on their explicit ratings, some rely on the similar age and gender groups, whereas some others rely on all features. According to the relevant experiments carried out in the study [Al-Shamri, et al., 2008], many cases indicate that, for different users, the feature weights vary greatly, and sometimes for a specific user, some features do not contribute at all during the recommendation process. Especially, when applying deep learning techniques to the system at scale, it is necessary to consider the efficiency of the models [Acun, et al., 2021]. Demonstrated in [Rungsuptaweekoon, et al., 2017] that adjusting the trade-off relationship of throughput and power efficiency is necessary for the deep learning applied system.

Optimising the features' weight individually for each user outperforms optimising the features' weight for all users together. However, tuning features' weight for each user individually is very computationally time-consuming [Al-Shamri, et al., 2008]. In other words, this weighting strategy pursues personalisation of the system by constructing a user model for each user; it discards the generalisation of conventional user model and sacrifices the computational efficiency of the system. Also, such weighting strategy might be very time sensitive. For an active user, some information about his/her profile might change very frequently when an interaction occurred with events such as rating new items or making new comments. This situation implies that, for active users, the user model needs to be updated frequently in order to maintain the recommendation accuracy. For the future research of micro learning, especially for the underlying fast-growing data, it is necessary to carefully balance the trade-off between the degree of recommender's personalisation degree and the workload of computation behind the system.

2.2 Preparation of the Online Learning Resources

As the micro learning service aims to contract a free and open learning environment, in some platforms, every registered user can create and upload new learning resources. The newly uploaded resources could be in an unstructured and informal format. These resources are hard to be directly

processed by the micro learning system for analysing or assessing purpose. Hence, it requires the micro learning system has an intelligent module that can automatically pre-process the continually generated new learning resources. Therein, extracting the key information and interpreting the resource content are two significant pre-processing steps.

2.2.1 Information Extraction

2.2.1.1 Background of Information Extraction

As a form of online service, micro learning targets massive online users of different ages, from different locations, and with different learning demands and preferences. Depending on the learners' learning purpose and the credit they can obtain after completing an entire course, this learning service could be formal, informal, or non-formal [Eshach, 2007]. The freedom and personalisation of this service further require the involved learning materials also to be diversified, which conform with the characteristic of 'big data' [Lin, et al., 2019]. However, operating and maintaining such an online service faces the difficulty of effectively managing dynamic and massive information from both the user and the resource side. In addition, sometimes, for a particular single application scenario, a large percentage of information is redundant or useless. Hence, effectively analysing information stream and precisely extracting valuable information become the significant and necessary pre-processing stage for the micro learning service. Such pre-processing stage can save budget, release the heavy workload, and improve the quality of the online service.

In a broad sense, information extraction refers to the task of automatically analysing, locating, distilling, summarising, and extracting useful information from massive unstructured multimedia documents. For the micro learning service, the utility of the information extraction technique could play a vital role in the pre-processing stage. The extracted information could be keywords/phrases, name entities, knowledge concepts, and many other types of key information, which will be used to train other models (like the recommendation model) of the system. Usually, for the online learning service, the raw information always contains temporal information. For example, historical learning records of an online learner contains interactions between the user and resources is in chronological order; the text content of a learning resource is also composed in sequential order. Hence, in this research, we mainly focus on investigating extracting valuable information from the sequential data.

2.2.1.2 Related Work of Information Extraction

In one previous work about cloud-based micro learning system [Sun, et al., 2018], the authors point out that keyword extraction is important to learning resource modelling, and real-time data extraction and inferencing is vital for constructing a personalised learner model. In another online learning related study [Sun, et al., 2018], the authors argued that extracting certain types of information like a user's profile is essential to solving the cold-start problem for recommending online educational resources. According to [Sun, et al., 2017], for an online learning service, intelligent components such as entity extraction, relationship extraction, and resources disambiguation are based on successfully extracting useful information from massive information stream.

Intuitively, the process of an information extraction task can be roughly divided into two individual steps: locate the valuable information from a given multimedia document and then classify the located information into the predefined categories. However, constructing two separate models will make the whole information extraction procedure error-prone. Especially for the micro learning service, the massive volume of information that needs to be extracted and labelled leads to the according to the rise of the complexity of their permutation and combination. It is necessary to reduce the error rate as low as possible. The better way is to design an end-to-end solution based on deep learning technology. An end-to-end solution which is capable to deal with two above-mentioned sub-tasks at the same time is vital to maintaining the robustness of the extracting procedure and effectiveness of the personalised service. Such information extraction solution takes input of the raw information and outputs the distilled valuable information, no intermediate output or operation involved. Moreover, the end-to-end model does not require many feature engineering tasks and domain knowledge [Guo, et al., 2017, Ma, et al., 2016], which makes it have satisfactory generalisation performance. Many prior studies suggested using the sequence labelling model for solving the information extraction problem. In a sequence labelling model for micro learning, the information stream is fed into the model in chronological order. The output is the sequence of relevant labels for data elements, each label can indicate the usefulness and category of each data element at the same time. Based on this idea, several representative prior studies about sequence labelling will be discussed in the following of this section.

Hidden Markov Model (HMM), Conditional Random Field (CRF), and their variants have been widely used in many previous studies for modelling various sequential and temporal problems

[McCallum, et al., 2000, Zhu, et al., 2005, Li, et al., 2018]. These models perform well in modelling local characteristics and constraints. However, such modelling strategies are based on linear information and heavily rely on handcraft features, which make them lack abilities to model complex tagging problems in a more general use case such as online non-formal learning.

Unlike other types of neural network, the recurrent neural network (RNN) shows outstanding ability in modelling the temporal dynamic information in many discipline areas. It can memorise historical information and combine such historical information with currently received information before making predictions. As one class of the RNN family, the Long Short-Term Memory (LSTM) based model shows satisfactory performance in modelling both short-term and long-term memory [Sutskever, et al., 2014] information. Moreover, bidirectional Long Short-Term Memory (Bi-LSTM) can further utilise 'future' input information for modelling sequential patterns [Dyer, et al., 2015]. Hence, the Bi-LSTM model can not only mine the forward sequential information but also mines the backward sequential information. However, such type of model does not perform well in modelling local constraints, especially when adjacent outputs in a sequence have a strong influence on each other.

Based on the advantages and disadvantages of various sequential modelling strategies, LSTM-CRF and its variants are state-of-art solutions for dealing with the sequence modelling problem. As pointed in a prior study [Huang, et al., 2015], such a network has the ability to efficiently use past input features from via LSTM layer and sequence level local information via a CRF layer. With the advantages of Bi-LSTM, the bidirectional LSTM-CRF (Bi-LSTM-CRF) model is used in many studies for different sequence tagging or labelling tasks [Ma, et al., 2016, Huang, et al., 2015, Yang, et al., 2018]; and in many cases, it outperforms the LSTM-CRF based model.

2.2.2 Text Analysis

2.2.2.1 Background of Multimodal Information Processing

With the development of Internet and communication technologies such as IoT and 5G, the forms of online information resources develop from single form (e.g., pure text or pure picture) to diversification (e.g., in a multimodal form). For example, a single online learning course could contain textual information (e.g., the discussion from students), audio-video information (e.g., the recorded lecture content), and images (e.g., the attached slides) [Chen, et al., 2019]. Alternatively, in

the area of e-commerce, a list of goods on Amazon or Taobao can have text, images and even short video for its description [Lynch, et al., 2016]. Hence, solely processing a single format of information is no longer adequate to fully understand the online resources. Such changes pose new challenges for the intelligent multimodal information processing tool towards various big data application areas, including e-learning, digital health, traffic information systems, etc.

However, the research on information processing and fusion for handling different information formats is still relatively isolated even though many efforts had been made independently or diversely. For example, there is a lack of a robust universal model, which can handle the processing task of different types of information (i.e., multimodal information). Hence, multiple models are generally involved in a processing system to handle and interpret multimodal information. In the prior work of managing the open educational resource in the medical discipline [Zhao, et al., 2016], three different models are used to separately handle text, video stream, and pictures. Ultimately, a universal and re-useable model or a single model only requiring little adjustment for different types of information, can greatly reduce the complexity of the overall system.

2.2.2.2 Related Work of Content Understanding in NLP and CV

With the outstanding performance in extracting different granularities of spatial information, the convolutional neural network (CNN) based models have become the mainstream solutions to various computer vision tasks [O'Mahony, et al., 2019]. To name a few, for the task of image classification, the AlexNet [Krizhevsky, et al., 2017], which consists of five convolutional layers, can distinguish one thousand different objectives. For the task of object detection, R-CNN [Girshick, et al., 2015], Fast R-CNN [Girshick, 2015] and Grid R-CNN [Lu, et al., 2019] are well-adopted region-based CNN with the idea of the bounding box to locate target objects. Moreover, besides recognising objects, the CNN-based model can also work as a tracker in the task of object tracking. One representative object tracking model is FCNT [Bertinetto, et al., 2016], which takes advantage of the feature map from the model VGG [Simonyan, et al., 2014]. Multi-hierarchical independent correlation filter is also used for visual tracking [Bai, et al., 2020]. As the network goes deep, different types of features (such as edges, shape, and so on) are extracted and modelled successively, but the network will face the challenge of vanishing gradients. Hence, the ResNet [He, et al., 2016], which consists of 152 layers, is designed to maintain the robustness of the deep model. All these works demonstrate the outstanding non-linear transforming ability of the CNN mechanism, which

can be used to interpret the content of the complex data or signals in the format of pixels.

In NLP, constrained by 'rules' (such as grammar and idioms), a meaningful sentence is commonly composed with specific patterns. When interpreting the meaning of a piece of text, a sequence of a bunch of words is usually more important than the individual words themselves. Due to the significance of modelling temporal/time-series information, the recurrent neural network (RNN) based model can be regarded as the dominant solution for various NLP tasks in recent years. For example, the combination of RNN and conditional random field (CRF) has been regarded as the optimal choice for the tasks of sequence labelling and information extraction in various application areas [Zhai, et al., 2018]. Similarly, for the task of text classification and text generation, most of the solutions are based on the famous RNN framework and its variants [Pawade, et al., 2018].

From the viewpoint of mathematics, the goals of both the CNN and RNN networks are to project the given input to the required output through complex non-linear transformation. Many prior studies have been successfully using the CNN model to solve NLP problems or the RNN model to solve CV problems in recent years. In [Zhai, et al., 2018], CNN is used to generate word embeddings by extracting character-level semantic information such as prefix and suffix. TextCNN [Kim, 2014] is designed for the task of sentence classification. This work demonstrates that a simple CNN with little hyperparameter tuning has the potential to achieve excellent results on multiple benchmarks. In the area of CV, the combination of CNN and RNN is used in [Wang, et al., 2016] and [Guo, et al., 2018] for image classification. The LSTM-C framework is used to incorporate the knowledge from external sources in another prior study [Yao, et al., 2017] t to address predicting novel objects in image captioning task. However, these studies do not dig and explore the relationship between these two areas in further depth and do not analyse whether it is feasible to have a universal framework or model for both areas.

2.3 Recommendation

As the key to personalisation, a recommender system, to a great extent, determines what kind of information will be finally delivered to the users. A good recommendation strategy should have the ability to automatically adjust the type of information to be delivered based on the user's background and the surrounding environment of the current learning activity. Compared to the other domains like e-commerce or entertainment, a recommending task in the educational domain has several unique

characteristics and requirements:

1. The learning activity and the learner profile always contain vague and uncertain information [Wu, et al., 2015]. A subject can belong to several different categories. For example, the subject 'statistical machine learning' is mainly relevant in the computer science area but also involves mathematics. Sometimes similar courses have totally different names such as 'Java' and 'Object-Oriented Programming'. And for a subject, it can have different difficulty levels for learners with different knowledge levels.
2. Pedagogical issues also influence recommending procedures significantly [Wu, et al., 2015]. Items liked by certain learners might not be pedagogically appropriate for them [Sikka, et al., 2012]. Unlike the recommender systems in the entertainment or social media domain, in the educational area, many subjects have various prerequisite courses. Also, for a certain learning period, a review or quizzes always need to be involved for pedagogical purposes.
3. As the micro-learning units vary in types (such as lecture, quiz, and tutorial) and formats (such as PDF, video, and audio) [Al-Hmouz, et al., 2012], recommending process should also consider how to choose the most suitable format and type of a learner based on the different contexts.

2.3.1 Recommender System for Online Learning Service

Recommender systems have been studied for many years based in various application areas. However, as mentioned above, due to some pedagogical considerations, a recommendation strategy used in other domains cannot be directly transformed to fulfil the requirements of online learning services. This is particularly true for the micro learning service. Hence, it still lacks a sophisticated solution to enable recommending personalised online resources to target users.

Ant colony optimisation (ACO) algorithm is used in many studies to tackle the path planning/recommendation problem. In [Zhao, et al., 2016], an ACO model is proposed to detect learners' learning transition such as knowledge area, and learning goal. In this study, the authors suggest that similar learning paths could represent a certain learning goal and learning requirements of a certain group of learners, and a learning path which is finished by a large number of learners could be seen as a valid or optimal learning sequence. This unsupervised learning path recommending strategy is self-adjusting and does not require labelled data. However, ACO is very

sensitive to the cold start problem as it requires predefined paths at the initial stage of the execution of algorithm. Interestingly, another study used improves ACO with adopted Mahalanobis distance to recommend learning paths to learners [Chen, et al., 2017]. Such model can avoid the side effects of the high dimensional space problem.

As another branch of soft computing, genetic algorithm (GA) has been widely used as an adaptive weighting method in the recommender system [Al-Shamri, et al., 2008, Fong, et al., 2008, Bobadilla, et al., 2011], which can further optimise the user model and boost the performance of collaborative filtering-based recommender system. Unlike conventional optimisation method, such as gradient descent, in GA, the searching for the optimal solution is mainly based on three genetic operators, namely selection, crossover, and mutation. The highlight of GA parameter tuning process is the operation of mutation and crossover, which can 'break the box' and find the new combinations of factors not being captured and recorded in the training dataset. With such ability, GA can explore new feature combinations in the global area with a relatively small sample set, it also alleviates the requirement of large amount data in the training step. When using GA to explore feature combination, some rare combinations of features could be explored as well, even if they would not appear in the training set. Different to the ant colony optimisation, which is mainly used for solving route optimisation like learning path design, particle swarm optimization (PSO) is mainly used for tuning parameters [Wasid, et al., 2015, Ujjin, et al., 2003]. The principle of PSO is to mimic the movement of an organism population, such as birds and bees; each individual has a trend moving to close to the 'optimal' point/area in their living environment based on the historical movement information of the whole population and itself. Similar to the GA as discussed above, PSO can improve the collaborative-filtering recommending result by well tuning the weights of the involved user model [Wasid, et al., 2015, Ujjin, et al., 2003]. Comparing to GA, PSO requires less computational time while demonstrating higher accuracy [Wasid, et al., 2015, Ujjin, et al., 2003].

Reinforcement learning (RL) has been widely used in various domains for sequential decision-making. The idea behind RL is to make sequential decisions (and take continuous actions) in an environment to maximise the notion of the cumulative reward. As mentioned above, learning path design is the extension of a recommender system in micro learning. With the concept of RL, a learning path can be naturally seen as a sequence of individual learning activities, and the knowledge gained from the learning activities can be seen as the accumulated rewards from the learning

activities. RL is utilised for choosing the difficulty level of learning materials for learners in [Fenza, et al., 2017]. In this study, the proposed recommending strategy was based on the Vygotsky's Zone of Proximal Development (ZPD) theory [Vygotsky, 1980]. ZPD is used to define and quantify the 'area' most suitable for a learner based on cognitive and affective perspectives. In ZPD, a learner would be kept in his/her leading edge which fell in between 'confused' and 'bored' status; this area challenged but would not overwhelm the learner [Murray, et al., 2002]. [Fenza, et al., 2017] also borrows the solution of the Cliff-Walking problem [Sutton, et al., 1998]. The goal of the Cliff-Walking problem is to find an optimal route, which can arrive at goal state 'G' from the starting stage 'P' or any other positions without stepping into the 'cliff' area.

A blended model is proposed in a prior study [Hoic-Bozic, et al., 2015], which combines a learning management system, a set of web 2.0 tools and the e-learning recommender system to enhance personalised online learning. However, this study does not provide technical innovations of a recommender system for online learning service. In an early survey [Sikka, et al., 2012], several recommendation approaches for e-learning service are listed and analysed. However, they are all too preliminary to be applied to the micro learning services. Another study [Chen, et al., 2014] proposes a hybrid recommendation algorithm which combines the collaborative filtering and sequential pattern mining together for a peer-to-peer learning environment. Learning path recommendation is investigated in [Rusak, 2017] and [Zhao, et al., 2016]. In [Zhao, et al., 2016], the ACO algorithm is proposed to recommend personalized learning paths to users based on the demographic information. The ontology-based method is used to add extra user's profile information for relieving the cold-start problem for micro learning service [Sun, et al., 2017, Sun, et al., 2018]. Study [Rusak, 2017] investigates the learning path recommendation for micro learning service from an exploitation perspective. However, the proposed models are constructed mainly based on demographic information, which does not provide much scope for exploring individual preferences. So far, there are few efforts on deep learning solutions to this problem.

A content-based convolutional neural network (CBCNN) recommender system is proposed in a prior study [Shu, et al., 2018], which shows fairly satisfying ability in mining new or unpopular learning materials for a target learner. Based on the usage of the online learning materials, another study proposes a new way to calculate similarities between online learning materials for recommendation tasks [Niemann, et al., 2013]; and the authors in that paper argue that the usage context-based model

has potential to outperform the content-based model, if the usage data is sufficiently fine-grained. And a system for recommending OERs in MOOC is proposed in [Hajri, et al., 2017], which has emphasised the significance of modelling users and learning materials. However, none of these studies mentioned the significance of the ranking for the success of an online learning service.

2.3.2 The state-of-the-art Recommendation Strategies

In many recommendation scenarios, the involved features have impacts on each other, aka feature interaction. For example, the feature pair (learning interests, knowledge level) determines the difficulty of a specific course for a learner. A classic work Factorization machine (FM) [Rendle, 2010] uses the inner product to model the feature interaction. This alleviates the data sparsity problem by using embeddings to represent the user and the item. However, due to computational and space complexity, only up to second-order feature interaction can be applied to many real-world applications.

The deep learning technique has been used in many application areas [He, et al., 2016, Fischer, et al., 2018, Liu, et al., 2017] and has demonstrated its outstanding ability to model complex problems. Hence, many researchers are investigating combining the deep learning technique with conventional recommendation strategies. One representative model proposed by Google is 'Wide&Deep'. This model combines the benefits of memorization and generalization by using a linear model and a deep network [Cheng, et al., 2016]. Both low-order and high-order feature interactions are also investigated in [Guo, et al., 2017] through FM and DNN components. As the DNN models high-order interactions in an implicit manner, the learned results can be arbitrary. In another work, an extreme deep factorization machine (xDeepFM) is proposed for generating and modelling feature interactions of a bounded degree [Lian, et al., 2018]. Similarly, model [Lian, et al., 2018] also contains two components, a compressed interaction network (CIN) and a DNN. The CIN and DNN learn the explicit and implicit feature interaction, respectively.

With the successes of 'Wide&Deep' and DeepFM [Guo, et al., 2017], the multi-component network structured recommender system becomes increasingly popular. Such structure shows outstanding performance in merging different techniques to mining latent information from different perspectives simultaneously. Among them, feature interaction and the weighting strategies are mostly benefitting from such type of multi-component structured network.

Feature interaction is a fundamental problem and plays a significant role in a recommendation task. There are also many prior studies [Wang, et al., 2017, Song, et al., 2019] mainly focusing on feature interaction strategies. The study [Wang, et al., 2017] proposes a novel cross-network, which explicitly applies feature interaction in each layer, and the cross-network consists of all the cross-terms of degree up to the highest. In another study [Song, et al., 2019], the key-value attention mechanism is used for determining which feature combinations are meaningful.

Attention mechanism has been widely used in many areas, such as computer vision and natural language processing. This allows the network to pay different degrees of attention to different parts. The attentional factorization machine is proposed in [Xiao, et al., 2017], which can distinguish the importance differences of various feature combinations. Instead of a simple attention network, the multi-head attention mechanism [Vaswani, et al., 2017] is also used in the study [Song, et al., 2019]. This study shows that the multi-head attention mechanism has the ability to explore meaningful feature combinations through different non-linear transformations. Squeeze-and-Excitation network (SENET) [Hu, et al., 2018] is used in the study of Feature Importance and Bilinear feature Interaction network (FiBiNET) [Huang, et al., 2019]. SENET is used to make the model pay more attention to the important features and decrease the weight of uninformative features by using the inner product and Hadamard product.

As demonstrated in [He, et al., 2016], the idea of the residual mechanism shows outstanding performance in stabilizing the optimization process of a deep network. Moreover, the residual function can also improve the model performance by providing sufficient information from previous layers of the network. Hence, for the recommendation task, as the network becomes deeper, many researchers start involving the residual connection (unit) in some components of the network. One prior study [Shan, et al., 2016] uses residual units to implicitly perform specific regularizations leading to better stability. Similarly, in the crossing component of the 'Deep&Cross' Network [Wang, et al., 2017], the residual unit is used in each crossing layer to add the current input information back. In [Song, et al., 2019], standard residual connections are added in both interaction and output layers to achieve a hierarchical operation manner.

2.3.3 Data Challenges for Research

When investigating the solution of a recommender system for the micro learning service, besides the

technical difficulties, there are also data challenges that exist.

2.3.3.1 Insufficient Data Source

The term ‘insufficient data’ means the data used in a study can only partially reflect the underlying issues against the context of potentially bigger data, and the experiment result of recommendation might be biased. Many current studies of e-learning recommender systems are based on a small amount of data.

The study [Chen, et al., 2014] proposed a hybrid recommendation algorithm which can reflect the timeliness of a learning procedure, but only 30 students are involved in this study. Metadata for Architectural Contents in Europe (MACE) and TravelWell datasets are used in the study [Niemann, et al., 2013] for training the usage context-boosted recommender system. However, both datasets contain very few users and only a fraction of subjects. One prior study [Shu, et al., 2018] uses convolutional neural network (CNN) to model the latent factors based on the Book-Crossing [Ziegler, et al., 2005] dataset. In this study [Shu, et al., 2018], the authors argue that CNN has the ability to better mining the textual information and can further boost a content-based recommender system. However, the type of learning materials could be video, audio, and text. Therefore, solely using a book dataset is insufficient for training a sophisticated recommender system for e-learning, especially for micro learning whose primary type of learning materials are in the video format. The fault-tolerant and the capability of self-optimization make Swarm Intelligence and Evolutionary Computing applied in many studies for learning path optimization [Zhao, et al., 2016, Chen, et al., 2017, Dwivedi, et al., 2018]. However, only 80 students are involved in the experiments of the study [Zhao, et al., 2016], and only one chapter of high school mathematics is used to train and validate the model in the study [Chen, et al., 2017]. Due to the heterogenetic between users and between subjects, insufficient subjects cannot generally reflect the underlying latent patterns or information of e-learning scenarios, and insufficient users cannot reflect the whole user population with different background and learning requirements under the big data era.

2.3.3.2 Inaccurate and Unknown Data Source

In addition, inaccurate and unknown datasets used in the research can impede the development of the research of micro learning. An inaccurate dataset refers to the dataset coming from noisy sources, other domains, or even simulated. An unknown dataset refers to the dataset used in the research, but

the researchers do not mention its source. Therefore, the experiment results obtained from such data set could be problematic, unconvincing or unrealistic.

Simulated data is used in the study to construct long short term memory (LSTM) model for learning path recommendation [Zhou, et al., 2018]. However, the simulation procedure and the validity of the simulated data are not mentioned in this study. The simulated data could be inaccurate, as in most cases we have no prior knowledge about users and learning materials such as their distribution. Due to lacking required dataset from the e-learning domain, another prior study [Wu, et al., 2015] uses MovieLens [Harper, et al., 2016] dataset to demonstrate that a fuzzy-tree-based collaborative filtering model has the potential to boost the recommending result. However, MovieLens is not a dataset from the educational domain. Using such an irrelevant dataset might not truly solve the recommendation problems in the e-learning domain.

Many studies about the recommending strategy are based on the applications or learning platform which developed by researchers themselves [Fenza, et al., 2017, Chen, et al., 2014]. For constructing the proposed recommending model, the study collects data from an educational game application called 'Itsego' and the feedbacks of expert teachers [Fenza, et al., 2017]. Similarly, the study [Chen, et al., 2014] is based on a learning system developed by researchers themselves. Most of the datasets from self-developed platforms or the LMSs hosted by researchers' affiliations are not open to the public. The study [Dwivedi, et al., 2018] collects four years of program from the CSED department in MNNIT Allahabad and uses the genetic algorithm to recommend learning paths. Prior research [Al-Hmouz, et al., 2012] demonstrates that an adaptive Neuro-Fuzzy inference system can recommend users with a suitable format of learning materials based on the users' current status and surrounding environments. Another study [Dorça, et al., 2017] uses clustering techniques to group learners with similar learning style together before recommending relevant learning objects. But these three studies do not claim the sources of the datasets they used and the acquisition procedure of the data sets. Such non-public datasets or the datasets with unknown sources make these studies difficult to be followed up, imitated, validated and further improved by other research groups.

2.4 Dropout Rate Prediction

2.4.1 Background of Dropout Rate Prediction

Since its first introduction in 2006 by Stanford University, learning through Massive Open Online

Courses (MOOC) has become one of the most popular online learning channels. Famous MOOC platforms, such as Coursera, EdX, and Udacity, provide high-quality learning resources to learners globally every day. So far, there have been more than one billion active MOOC users around the world. Prior to the COVID-19 pandemic, as of 2019 with a confirmed record, more than nine hundred universities have offered online courses on various MOOC platforms. Despite this advancement, a high dropout rate for MOOC courses persists [Kloft, et al., 2014]. The causes for dropout rates have been difficult to precisely pin down or predict. It could be the quality issue of the quiz question, or the unsatisfactory presentation of the course material, or maybe even just some personal issues. At the early stage of a learning activity, the ability to accurately and timely identify the "at-risk" student with a high probability of dropping out is vital for the continuous development of an efficient and intelligent MOOC platform and online learning. Also, as highlighted in the prior research, dropout prediction can be an effective precursor for early intervention [Whitehill, et al., 2015]; it can also be used as a pedagogy enhancement method to decide whether an online course needs adjustment or modification [Tang, et al., 2018].

In general, an online course is composed of several instructional videos and each video contains a certain number of complete knowledge points. Such course structure can offer two types of information for the dropout prediction task, micro information and macro information. The coarse-grained macro information refers to the overall profile of a course, like a discipline area it belongs to and the difficulty level of the course. The micro information is fine-grained information, which refers to the interaction detail between a user and a certain video, such as duration and the start time of video watching. Both types of information offer valuable information about the course itself and how a user interacted with this course historically. Accurate dropout rate prediction depends on whether a prediction model can robustly incorporate and interpret these two types of information. However, to the best of our knowledge, the research about how to properly handle these two types of information still remains less touched.

2.4.2 Related Work of Dropout Rate Prediction

With growing interest in applying the machine learning technique to the e-learning problems in the multi-disciplinary research community, learner dropout prediction models have been generated using various techniques [Wang, et al., 2017, Dalipi, et al., 2018]. They include logistic regression and its

variants [He, et al., 2015], support vector machine [Kloft, et al., 2014, Amnueypornsakul, et al., 2014], and decision trees [Al-Shabandar, et al., 2017, Al-Shabandar, et al., 2017]. Deep learning-based models have also shown great potential in mining complex latent information for the task of dropout prediction in more recent years. For example, multi-layer perceptron (MLP) with different combinations of hidden layer architectures for dropout prediction is used with an in-depth comparison concerning various evaluation metrics [Imran, et al., 2019]. Model ConRec is proposed by [Wang, et al., 2017], which combines the merits of the recurrent neural network (RNN) and the convolutional neural network (CNN). A context-aware feature interaction model is proposed in [Feng, et al., 2019], which utilises the technique of context-smoothing to smooth feature with different contexts and combine user and course information by the attention mechanism. However, to the best of our knowledge, for the task of learner dropout prediction, none of the prior studies have tried to distinguish and model the micro and macro information separately; this has primarily motivated this paper.

Benefiting from its effectiveness in extracting complex information sequences, the deep learning-based model has become the preferred choice for modelling temporal information. Typical examples include using bidirectional long short term memory (Bi-LSTM) based network for labelling visual sequence [Koller, et al., 2017]. For name entity extraction in natural language processing (NLP), the combination of LSTM and conditional random field (CRF) and its variants have been adopted in many applications [Huang, et al., 2015]. For monitoring machine health, some authors proposed a local feature-based Gated Recurrent Unit (GRU) [Zhao, et al., 2017]. Similarly, the position-aware bidirectional attention network (PBAN), based on GRU, is used in [Gu, et al., 2018] to distinguish each specific aspect's sentiment polarity in a given sentence. In [Kim, 2014], as one of the earliest efforts, the authors demonstrate that the convolutional neural network (CNN) could also be used to model time-series information.

However, for different application scenarios, the characteristics of the time-series information vary and are modelled differently by using various architectures of deep learning networks. Hence, directly using the mainstream solutions (like LSTM-based network) might not be the optimal solution for the task of dropout rate prediction. To date, there is little effort in comparing the efficacy of different time-series modelling techniques in predicting the learner dropout rate, as required for this research.

2.5 GAN Framework for the Micro Learning Recommender System

2.5.1 GAN and Its Milestones

The idea of GAN [Goodfellow, et al., 2014] is firstly proposed in 2014. Until then, the mainstream generative models are Restricted Boltzmann machine (RBM) [Smolensky, 1986], deep belief network (DBN) [Hinton, et al., 2006], and Autoencoder [Kramer, 1991]. Comparing with RBM, DBN, and Autoencoder, GAN shows the advantage of higher flexibility. GAN can be conditioned by different input with diverse network structures. The original GAN contains two networks, a generator G and a discriminator D, each can be a simple multilayer perceptron. The framework structure of GAN is shown in Figure 2.1. The generator produces fake sample to fool the discriminator, whereas the discriminator differentiates the fake sample from the real sample.

The training process of the GAN is based on a minimax gaming algorithm in which a generator and a discriminator compete against each other until an optimal point (i.e., the Nash Equilibrium [Heusel, et al., 2017]) is reached [Torres-Reyes, et al., 2019]. The objective function of this process can be formulated as Equation (2.1):

$$\min_{G} \max_{D} V(D,G) = \mathbb{E}_{x \sim P_{data(x)}}[Log(D(x))] + \mathbb{E}_{z \sim P_z}[Log(1 - D(G(z)))] \quad (2.1)$$

x is the real data over the distribution of P_x, z is the input noise over the distribution of P_z. $D(x)$ denotes the probability that x is the real data. $Log(D(x))$ and $Log(1 - D(G(z)))$ are the cross-entropies.

Generally, there are two schools of thoughts competing to further develop the GAN technique. The first is application-oriented and focuses on incorporating the state-of-the-art machine learning techniques into the GAN framework and applying them in various real-life tasks such as image generation and text generation. The second is theory-oriented and is motivated by the view that GAN is still intricate to implement and difficult to tame. Therefore, this second school of thought focuses on working out its training, evaluating, and optimising problems from the theoretical viewpoint to ameliorate existing drawbacks of GAN.

GAN and its variants have been applied in various application areas. Herein we mainly focus on representative first attempts and milestones about applying the idea of GAN in a new domain.

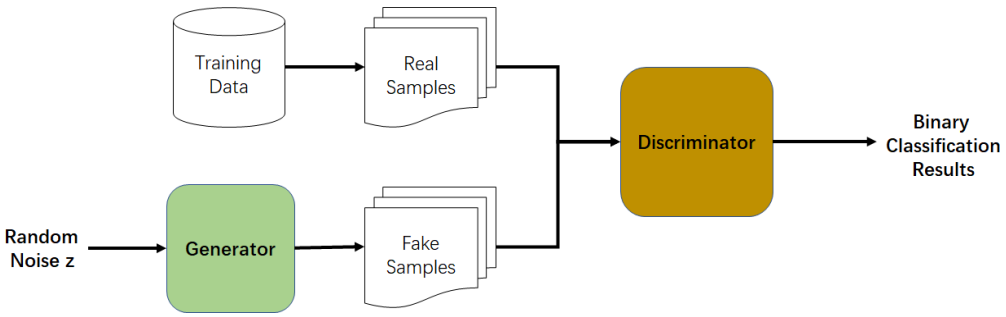

Figure 2.1 Structure of the GAN Framework

Using adversarial training to generate text sequence has been most often discussed. TextGAN demonstrates that GAN can produce realistic sentences by mimicking a real input sentence, and the learned latent representation space can continuously encode plausible sentences [Zhang, et al., 2016]. The study of SeqGAN [Yu, et al., 2017] integrates sequence modelling strategy with the GAN framework. This is the first work extending GANs to generate sequences of discrete tokens. It is very informative in signal processing and sequence modelling, such as text generation and music generation. IRGAN [Wang, et al., 2017] is also a big milestone in information retrieval. Indeed, it is the first study introducing the GAN technique to the recommendation problem. Details of this study will be later reviewed in Section 2.5.2.

DCGAN [Radford, et al., 2015] primarily implements the GAN framework with a convolutional neural network. This study also discusses and analyses the feature visualisation of the GAN framework and the interpolation problem of latent feature space. Considered as a huge leap in computer vision, vid2vid is capable of synthesising 2K resolution video up to 30 seconds long [Wang, et al., 2018]. As the extension of the prior work [Wang, et al., 2018], the vid2vid involves extra information about optical flow in the generator and the discriminator, and constructs models for foreground and background. StyleGAN [Karras, et al., 2019] is based on an alternative generator architecture. This model can be regarded as one of the most complicated GAN models in recent years. This work uses adaptive instance normalisation (AdaIN) to control the vectors in the latent space. This work has demonstrated that StyleGAN has a surprising ability to automatically learn, unsupervised separation of high-level attributes, and generated images entitled with a stochastic variation. In the task of infrared small object segmentation, GAN is used to balance miss-detection and false alarm. For example, one research demonstrates that, besides the powerful ability in generating the true-to-nature data [Wang, et al., 2019], GAN can be trained to form a sophisticated

classifier with proper adjustments (such as tuning the loss function and the structure of the GAN framework).

2.5.2 The Pioneer GAN solution for General Information Retrieval Task

Prior to the idea of using GAN in a recommender system, there are two extant information retrieval approaches, the generation process and the discrimination process. The classic recommender system assumes that there is an underlying generation process between items and information needs, $q \rightarrow D$, that the relevant documents D are generated/clued by a query q [Lafferty, et al., 2003]. For example, when processing the query "*weather tomorrow newyork*", a model aims to obtain documents clued by this query, ideally the one containing information about "*What's the weather like in New York tomorrow*". The modern approach argues that the process of the information retrieval problem is a discriminative task, no matter whether or not a document is relevant to the query, such as the representative work reported in [Koren, et al., 2009]. Under this assumption, queries and documents are joint as a single feature and their relevancy ($q+D \rightarrow r$) is provided. Thus, the model depicts the above example as: how the relevant degree of the query "*weather tomorrow newyork*" is, and the document contains the information "*What's the weather like in New York tomorrow*".

IRGAN combines the ground-breaking idea of generation and discrimination together by involving a generator and a discriminator. The generative model tries to generate or select relevant document based on the given query q; its goal is to approximate the true relevance distribution over the documents [Wang, et al., 2017]. While the discriminative model in IRGAN tries to discriminate well-matched query-document pair and ill-matched ones, which is determined by the relevance of a document and a query; the goal of the discriminator is to distinguish relevant documents and non-relevant ones for a given query. During the training process of IRGAN, the generative model is guided by the signal obtained from the discriminative model that minimises the probability that the discriminator can distinguish the real and generated samples. In contrast, the discriminative model is enhanced to rank top recommendations better by maximizing the log-likelihood of correctly distinguishing the real and generated samples [Wang, et al., 2017]. The objective function of the generator and discriminator are formulated as Equation (2.2) ~ (2.4).

$$J^G = max \sum_{i=1}^{N} \left(\mathbb{E}_{d \sim p_\theta(d|q_n, r)} \left[\log \left(1 + \exp \left(f_\phi(\mathrm{d}, q_n) \right) \right) \right] \right) \tag{2.2}$$

$$J^D = max \sum_{i=1}^{N} (\mathbb{E}_{d \sim p_{ture}(d|q_n, r)} \left[\log \left(\sigma \left(f_\phi(d, q_n) \right) \right) \right] + \mathbb{E}_{d \sim p_{\theta^*}(d|q_n, r)} [\log(1 -$$

$$\sigma(f_\phi(d, q_n))\big)\big]) \tag{2.3}$$

$$\sigma(f_\phi(d, q)) = \frac{\exp(f_\phi(d,q))}{1+\exp(f_\phi(d,q))} \tag{2.4}$$

Based on the pioneering work of IRGAN, the development of GAN-based recommender systems can be roughly divided into two categories. One focuses on investigating the end-to-end recommendation strategy, and another branch focuses on exploring how to augment data quality and boost the recommendation results. The representative works of both categories will be discussed in the following subsections.

2.5.3 The GAN-based Recommender System

This section will first show how GAN-based recommender systems have been applied to solve different recommendation problems and what distinct advantages these models have. The mathematic detail of the objective functions and the descriptions of the relevant concepts of these models are shown in the Table 2.1 and Table 2.2.

Table 2.1 The Details of the Objective Function Used in Reviewed Studies

Model	Object Function for G and D	Notes
CFGAN [Chae, et al., 2018]	$J^G = min \sum_u (log\big(1 - D(\widehat{\boldsymbol{r}_u} \odot (\boldsymbol{e}_u + \boldsymbol{k}_u) \mid \boldsymbol{c}_u)\big) + \alpha \cdot \sum_j (x_{uj} - \hat{x}_{uj})^2)$ $J^D = max \sum_u (log\, D(\boldsymbol{r}_u \mid \boldsymbol{c}_u) + log\,(1 - D(\widehat{\boldsymbol{r}_u} \odot (\boldsymbol{e}_u + \boldsymbol{k}_u) \mid \boldsymbol{c}_u)))$	Information about non-purchased item is involved in the loss functions of the generator and discriminator through zero-reconstruction term $\sum_j (x_{uj} - \hat{x}_{uj})^2$ and partial-masking ($\boldsymbol{k}_u$ vector denotes the negative items).
RecGAN [Bharadhwaj, et al., 2018]	$J^G = min \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{M} (\mathbb{E}_{r \sim D(r \mid i,j)_{\theta \mid t}}[\log (1 - D(r \mid i,j,t))])$ $J^D = max \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{M} (\mathbb{E}_{r \sim D(r \mid i,j)_{real \mid t}}[\log D(r \mid i,j,t)] + \mathbb{E}_{r \sim D(r \mid i,j)_{\theta^* \mid t}}[\log (1 - D(r \mid i,j,t))])$ $J^D = \max_{\eta} \frac{1}{\mathcal{T}_u} \sum_{T \in \mathcal{T}_u} (log \mathcal{D}_\eta(u,T) + \log (1 - \mathcal{D}_\eta(u, \mathcal{T}_{\mathcal{G}})))$	Temporal information is encoded into the loss functions by involving additional time indexes T. Both network of the generator and discriminator are constructed using Gated Recurrent Unit.

Real-value vector-wise adversarial training is proposed in CFGAN [Chae, et al., 2018], which shows the ability to fully exploit the advantage of adversarial training with higher accuracy in the prediction results. Moreover, three different reconstructed loss functions are proposed and compared, making the model not only focus on the positive user feedback but also the negative ones. The combination of recurrent neural network (RNN) and GAN technique is proposed in [Bharadhwaj, et al., 2018], RecGAN uses the customised Gated Recurrent Unit (GRU) to replace the original fully connected layer of the generator and the discriminator. Such modification makes the proposed model have the ability to capture temporal information from historical interactions between the users and the items [Bharadhwaj, et al., 2018].

Table 2.4 The Details of the Objective Function Used in Reviewed Studies (Continue)

Model	Object Function for G and D	Notes
KTGAN [Yang, et al., 2018]	$J^G = max \sum_{i=1}^{N} \mathbb{E}_{m\sim p_\theta(m\|u_i,r)}[\log(1 + exp(f_\phi(u_i,m)))])$	Side information are used through the score function $s(u,m) = \boldsymbol{e}_u \cdot \boldsymbol{e}_m + b_m$, and the softmax function is formulated as: $p_\theta(m_k\|u,r) = \exp(s_\theta(u,m_k)) \Big/ \sum_m \exp(s$
	$J^D = max \sum_{i=1}^{N} (\mathbb{E}_{m\sim real(m\|u_i,r)}\left[\log\sigma\left(f_\phi(u_i,m)\right)\right] + \mathbb{E}_{m\sim p_\theta(m\|u_i,r)}[\log(1-\sigma(f_\phi(u_i,m)))])$	
APOIR [Zhou, et al., 2019]	$J^G = min \sum_{u_i\in\mathcal{U}} \mathbb{E}_{l^R\sim R_{\theta(l^R\|u_i)}}\left[\log(1-\sigma(D_\phi(u_i,l^R)))\right]$	The combination of GRU and Matrix Factorization is used to model the temporal information of POIs. $l^R\sim R_{\theta^*(l^R\|u_i)}$ is the generated POIs by the current optimal R_θ, $l^+\sim\mathcal{L}^{u_i}$ are positive samples.
	$J^D = max \sum_{u_i\in\mathcal{U}} (\mathbb{E}_{l^+\sim\mathcal{L}^{u_i}}\left[\log D_\phi(u_i,l^+)\right] + \mathbb{E}_{l^R\sim R_{\theta^*(l^R\|u_i)}}\left[\log(1-D_\phi(u_i,l^R))\right])$	
LambdaGAN [Wang, et al., 2019]	$J^G = min \sum_{i=1}^{N} (\mathbb{E}_{o'\sim p_\theta(\boldsymbol{o}'\|u_i)}[1 - f(\mathrm{D}(\boldsymbol{o}',u),\varsigma_u)])$	ς_u is the item ranking list for user u, the ranking information is calculated through NDCG, which is formulated as $f(D(<i_p,i_q>\|u),\varsigma_u) = D(<i_p,i_q>\|u)\|\Delta NDCG_{p,q}\|$
	$J^D = max \sum_{i=1}^{N} (\mathbb{E}_{o\sim real(\boldsymbol{o}\|u_i)}[\mathrm{f}(\mathrm{D}(\boldsymbol{o},u),\varsigma_u)] + \mathbb{E}_{\boldsymbol{o}'\sim p_\theta(o'\|u_i)}[1 - f(\mathrm{D}(\boldsymbol{o}',u),\varsigma_u)])$	
DP-GAN [Wu, et al., 2019]	$J^G = \min_\theta \sum_{i=1}^{N} \log(1-\mathcal{D}_\eta(u,\mathcal{G}_\theta))$	DPP is used to sequentially sample top-K diverse items with kernel matrix L, the sampling process is denoted as $S_{K-DPP}(\boldsymbol{L})\sim\mathcal{G}_\theta(u)$. The reward function reflects how much the generator can deceive the discriminator is formulated as $\mathcal{R}(u,\mathcal{T}_\mathcal{G}) = -\log(1-\mathcal{D}_\eta(u,\mathcal{T}_\mathcal{G}))$.
	$J^D = \max_\eta \frac{1}{\mathcal{T}_u} \sum_{T\in\mathcal{T}_u} (log\mathcal{D}_\eta(u,T) + \log(1 - \mathcal{D}_\eta(u,\mathcal{T}_\mathcal{G})))$	

The framework proposed in the study of KTGAN illustrates how to incorporate the various pre-trained embedding results with a GAN-based recommender system [Yang, et al., 2018]. This framework makes the original GAN-based recommender system more extendable and flexible when involving different types of side-information based on the requirements of different tasks. Specifically, by integrating geographical and social information into the reward function, APOIR

obtains a significant improvement in handling the recommendation task of point-of-interest (POI) [Zhou, et al., 2019]. Two novel loss functions are used in the study of LambdaGAN [Wang, et al., 2019], which focus on making GAN-based recommender system gain the ability to precisely rank the recommendations by involving NDCG metrics [Burges, et al., 2007]. This study is the first study combining the lambda strategy with generative adversarial learning; its experimental result shows that the LambdaGAN outperforms the original IRGAN in pairwise scenarios with dataset MovieLens-100K[3] and Netflix[4].

How to balance the trade-off between relevance and diversity of the recommendation results is discussed and analysed in the research on DP-GAN [Wu, et al., 2019]. The pre-trained Determinantal Point Process (DPP) model [Chen, et al., 2018] is incorporated with the conventional GAN framework for capturing the diversity of items. Such improvement makes the generator to be able to generate a set of diverse and relevant items that are similar to the ground truths in order to fool the discriminator [Wu, et al., 2019].

2.5.4 The GAN-based Data Augmentation

Another branch of the studies focuses on utilising GAN to augment various kinds of information involved in the recommendation task. The representative workflow of GAN-based data augmentation for a recommender system is shown in Figure 2.2. The generated fake data and the real data are combined as the input of the recommender system during the training process. As the extension of CFGAN, [Chae, et al., 2019] proposes the RAGANBT model to augment rating information. The goal of the proposed model is to alleviate the data sparsity problem in collaborative filtering. Similarly, one prior study [Wang, et al., 2019] proposes AugCF based on the Conditional Generative Adversarial Net [Mirza, et al., 2014]. This model utilises the side information (e.g., the user's age and gender) to generate new interactions for less active users. In another study [Gao, et al., 2019], the authors argued that many datasets contain a huge amount of negative user feedback, but the quality of such feedback is low. Hence, the model DRCGR is proposed to generate high-quality negative samples from implicit user feedback (i.e., skipping behaviour during the interaction) [Gao, et al., 2019]. The generated negative items which a given user might not be interested in are produced in the study of Collaborative Adversarial Autoencoder [Chae, et al., 2019]. In this study,

[3] https://grouplens.org/datasets/movielens/
[4] https://www.kaggle.com/netflix-inc/netflix-prize-data

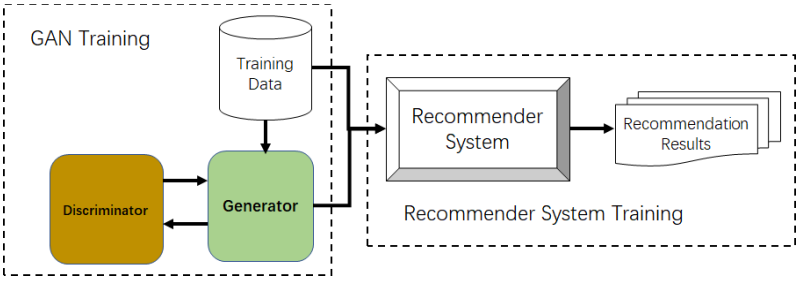

Figure 2.2 GAN-based Data Augmentation for a Recommender System

the autoencoder is used as the generator and the Bayesian personalised ranking model [Rendle, et al., 2012] is used as the discriminator.

Authors of [Sun, et al., 2020] propose LARA to generate fake user representations for the new item, and the similarity measurement is used to find the most similar users for the new item. This model aims to address the cold-start problem for new items by jointly modelling and obtaining the user-item attribute-level interaction information. A creative study about fashion recommendation uses GAN to generate complementary images to promote people's interest and participation in the online retail market [Kumar, et al., 2019]. The generated images could be further used to fetch purchasable items from any image search on an e-commerce platform.

GAN-based recommender systems are also proposed in several recent studies to solve cross-domain recommendation problem. In the work of CnGAN [Perera, et al., 2019], the user's preference information is generated by using the Twitter dataset and further used to solve video recommendation in the YouTube platform. These generated user preferences improve the results of recommendation for non-overlapped users. On the other hand, RecSys-DAN addresses the cross-domain data imbalance issue by adopting an adversarial loss for the cold start problem [Wang, et al., 2019].

2.6 Summary

In this chapter, the related prior works of micro learning and its required processing methods are discussed and analysed. In summary, the drawbacks and technical requirements of existing research can be concluded as:

1. Different application scenarios or domain require different pre-processing models. The pre-processing model for the specific application background of micro learning is not fully explored and comprehensively investigated by the prior research from the related area.

2. Little research has investigated the recommender system for micro learning service. Due to the domain differences, current the state-of-the-art recommender systems might not be suitable for the micro learning service. It is required to design a recommender system for the micro learning service basing on various existing recommendation strategies.
3. For educational purposes, a mature online learning platform/system requires to have some assistant plugins which can evaluate the course quality or assess the engagement of an online user. The dropout rate prediction is one way to reflect the current course quality or a student's future engagement. How to design a dropout rate prediction for the online learning service such as micro learning is vital for the development of online learning/micro learning.
4. And from the above reviewed studies, we can see that the GAN technique has demonstrated its advantages of generalisation and flexibility when dealing with complex real-world problems. Such characteristics of generalisation and flexibility are required for the recommender system of a micro learning service. However, how to apply GAN to this service is still not touch by researchers from either educational domain or computer science domain. We hold that the GAN-based model has the potential to further exploring latent information from user-item interaction logs for the recommender system of the micro learning service.

It can be concluded that, the research of the specific AI solution for many micro learning processing stage is still less touched. In most case, the researchers have to reference the solutions from other application domains. In this research, I investigate applying AI solutions to the pre-processing, recommendation, and for assistant purposes of the micro learning service. The following chapters of this thesis will show the AI solution to these tasks and problems.

Chapter 3 The Big Picture of the Micro Learning System

In this chapter, I propose the framework of the A.I. based micro learning system. Despite introducing each intelligent module, the dataflow of the system and the characteristics of the involved data sources are also discussed in this chapter.

3.1 System Framework

As a comprehensive intelligent system for the online learning service, the proposed micro learning system should contain a complete path of end-to-end resource/information processing chain. From a high-level point of view, the proposed system can be regarded as a bridge that connects the users and online learning resources (as shown in Figure 3.1). The learning resources are delivered to the target user through the micro learning system, and this procedure is driven by several intelligent modules of the system. In the meantime, the interaction records between the users and the learning resources are fed back to the system to further optimise the involved intelligent modules.

From the resource-side perspective, the processing workflow of the micro learning system is shown in Figure 3.2. As micro learning can be informal, the new learning resources are created from time to time by different online users. The newly created learning resources contain unregulated or unstructured information. Hence, the pre-processing module of the micro learning system takes in new learning resources and makes them ready for delivery or assessment. The recommender system takes in the pre-processed learning resources and ranks them. The ranked online learning resources

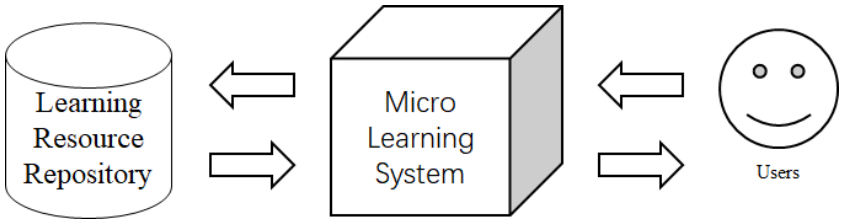

Figure 3.1 High-level Framework of the Intelligent Micro Learning System

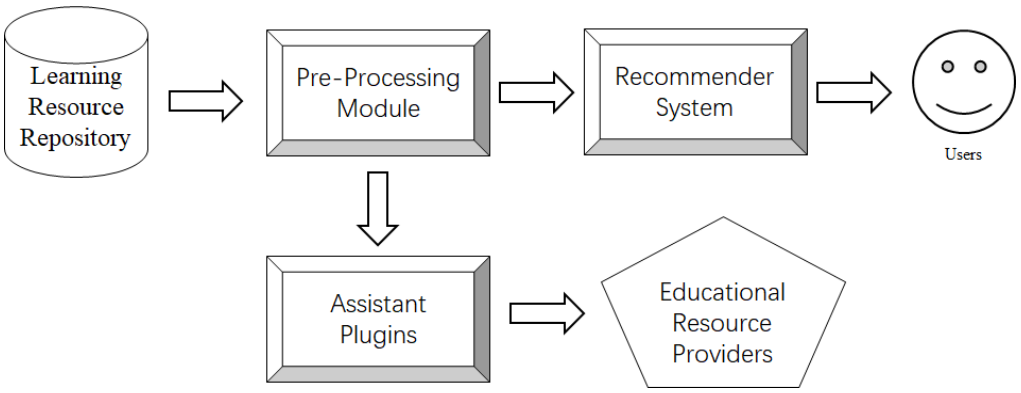

Figure 3.2 System Workflow (from the resource perspective)

are delivered to the target user during the learning activities. In the meantime, the information of the pre-processed learning resources is also passed into several assistant plugins for different evaluation/assessment purposes. These plugins can analyse the learning materials based on the pre-process information such as the course summary and extracted keywords/sentence of the course content. The final evaluation/assessment results will be fed back to the educational resource providers.

From the user-side perspective, the processing workflow of a micro learning system is shown in Figure 3.3. During the online learning activities, a user interacts with one or many learning resources. These interactions records are stored in the form of interaction logs. The interaction logs are fed back to the recommender system and the assistant plugins for the further optimisation of these modules. These logs can provide and reflect user-side information such as learning behaviours, which is significant for personalised online learning service. Similarly, together with the information provided

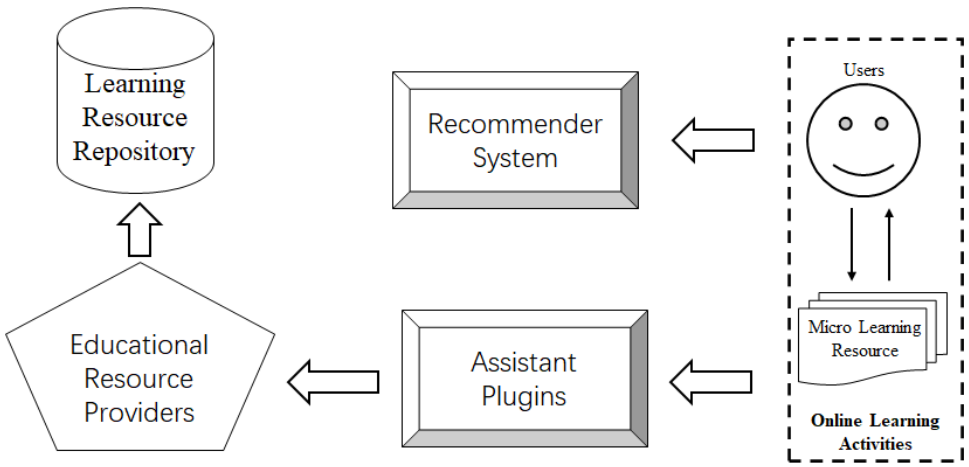

Figure 3.3 System Workflow (from the user perspective)

by a pre-processing module that we mentioned above, the assessment plugins can generate insight useful assessment information to the educational resource providers. Based on these feedbacks the providers can further update the learning resources or create better new learning resources.

3.2 Data Sources

3.2.1 The Utilities of Different Types of Data

A micro learning system consists of two core parts and several prediction and analysis modules, which work together to provide a complete real-time personalised online learning service. The summarisation of the utilities of the different types of data is shown in table 3.1. As discussed in the previous sections, most intelligent models involved in the system are driven by data. For a micro learning system, based on the data types, the required data source can be roughly classified into four categories: user' historical learning and interaction records from log files, users' profile and items' content information stored in the relevant databases, and other contextual information captured by the platform or client during the learning activities. Usually, the interaction records between the user and online learning resource are used to construct and optimise the recommender system. The content information of the learning resource is regarded as the raw material for constructing the

Table 3.1 The Utility of Different Types of Data

<table>
<tr><th colspan="3">Data Type</th><th>Utility and Description</th></tr>
<tr><td rowspan="4">User's Interaction</td><td rowspan="3">Learning Interactions</td><td>Clickstream</td><td rowspan="4">The interaction records between users and resources are indispensable for constructing a recommender system. The interaction record is also the main information sources for the tasks of performance prediction and analysis. This type of information reflects how users react in different learning activities. Various prediction and analysis models, such as dropout prediction and learning path design, are also based on such information. Demographic information (such as the popularity of a course) extracted from users' interaction can also be used to improve the recommendation.</td></tr>
<tr><td>Comments</td></tr>
<tr><td>User's Access log</td></tr>
<tr><td>Temporal Type Information</td></tr>
<tr><td colspan="2">Quiz/Exam Performance</td></tr>
</table>

resource profile. It is the main source to provide resource-side information for various kinds of task. The user's personal profile is another type of user-side information, which is usually provided by user himself/herself. Hence, it explicitly reflects user's subjective personal characteristics. Together with user's activity record, the user's profile is usually used for constructing the user model for the personalised online service. Lastly, the contextual information, which is usually captured in real-time, reflects the contextual characters of current learning activities. It is usually used for providing supplementary evidence for an intelligent model to further enhance the personalisation level of an online service.

Table 3.1 The Utility of Different Types of Data (Continue)

<table>
<tr><th colspan="2">Data Type</th><th>Utility and Description</th></tr>
<tr><td rowspan="3">Content Information</td><td>Textual Information (such as course title and description)</td><td rowspan="3">The Content information usually comes from the learning resource itself. The useful information is extracted in the pre-processing module of the system. The fine-grained content information is usually used for constructing the profile for the learning resource.</td></tr>
<tr><td>Audio/Video Information</td></tr>
<tr><td>Other Metadata (such as instructors, pre required knowledge, and difficulty level)</td></tr>
<tr><td colspan="2">User's Profile</td><td>The main source of the information about user explicit characteristics. This type of information is usually used for providing user's personal details for constructing the user model. The constructed user model can be used for the downstream personalized modelling purposes.</td></tr>
<tr><td colspan="2">Contextual Information</td><td>For a personalized online real-time service, the contextual information is usually used as the supplementary information in a decision-making process, which could be time, location, internet band width, or anything included in the learning activity. This type of information reflects users' real-time status.</td></tr>
</table>

3.2.2 The Data Flow and Characteristics

In the proposed micro learning system, different types of information exchanges among different intelligent models. The overall data flow in a micro learning system is shown in Figure 3.4. Based on the different characteristics of the above mentioned four types of data, they can be further classified into two categories: the dynamic and the static data.

3.2.2.1 Static Data

The content of the online learning resource and the user profile belongs to the category of static data. The static data means this type of data rarely changes after it is stored in the relevant database. For example, the content information could be the video content of a certain course, the course name, and the instructor of this video. Once a learning material is uploaded to the micro learning system, usually, its content and related information will not change significantly over a certain period of time. Hence, the content information of the learning resource is relatively static data. Similarly, in many online systems, the user will provide some basic profile when creating an account or the first time they log in. For an online learning system, the user profile usually contains the user's basic information such as gender, age, occupation, and learning preference. Such information is also relatively static and will not change often over a certain period of time. In a micro learning system, both content information and user profile are stored in the related databases. The content of the online learning resource and the user profile belong to the category of static data. The static data means this type of data rarely changes after it is stored in the relevant database. For example, the content information could be the video content of a certain course, the course name, and the

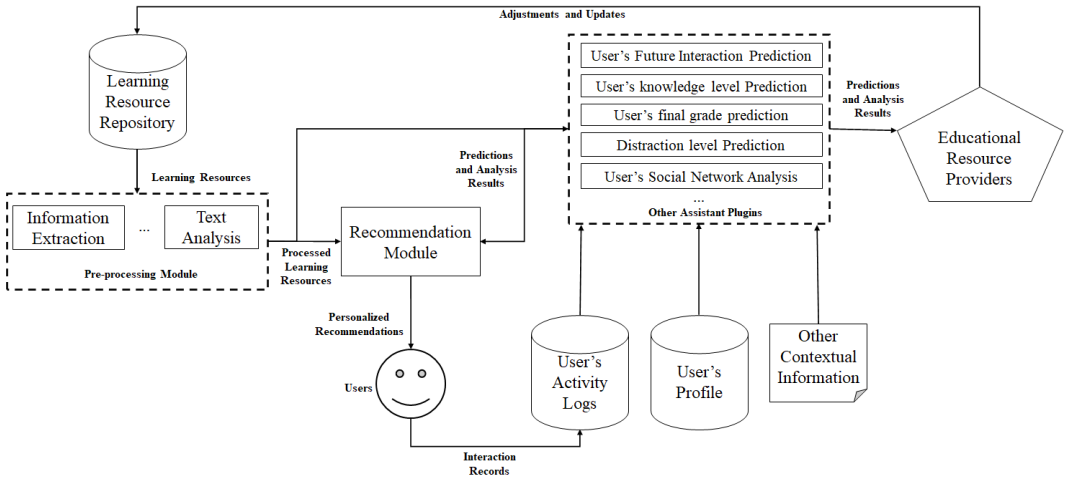

Figure 3.4 The Data Flow Detail of the Proposed System

instructor of this video. Once a learning material is uploaded to the micro learning system, usually, its content and related information will not change significantly over a certain period of time. Hence, the content information of the learning resource is relatively static data. Similarly, in many online systems, the user will provide some basic profile when creating an account or the first time they log in. For an online learning system, the user profile usually contains the user's basic information such as gender, age, occupation, and learning preference. Such information is also relatively static and will not change often over a certain period of time.

3.2.2.2 Dynamic Data

Dynamic data means this type of data changes or evolves frequently from time to time. As discussed in the previous section, contextual information is about the user's current environmental surroundings during a learning activity. Usually, the contextual information (such as geolocation and internet bandwidth) is capture in real-time by the client or system. Hence, it is relatively dynamic; and due to its time-sensitive characteristic (previous contextual information is not very useful for current learning activity), it is not stored in the databases of the system or only stored for a short time. The interaction record between the user and the learning resources is another type of dynamic data with incremental character. When a user interacts with a learning material each time, the system will generate a log file that records the interaction details such as the clickstream and date-time. The generated records will be processed and stored in the database for a certain period of time. Usually, the active user will have rich interaction records and the inactive user will have little such type of information.

3.2.2.3 Data Flow

Before the commencement of recommendation, there is a pre-processing stage to get micro learning materials settled. This part mainly focuses on dealing with the content of the online learning resource (content information). The content information of the learning resource is interpreted, analysed, and summarised in this pre-processing step. Most types of content information (course title, course description, and instructor) of the learning material are open to the public, while the demographic information (like the regional distribution of the enrolled students) is not. The pre-processing module takes the raw content information of the learning resource as the input. Then different A.I. models simultaneously process the raw input information and output the cleaned fine-grained information.

For the recommender system of a micro learning system, users' historical learning activities are indispensable for constructing and optimising the models, no matter what recommending strategy is applied. In most cases, the information about the user's historical activities only exists in the log files and cannot be crawled from online learning platforms or websites. As discussed in one prior study [Shu, et al., 2018], historical data-based recommendation methods require extensive historical data, which is difficult to obtain from the e-learning system directly. And with only the user's historical learning activities is not enough to fully describe the learning requirement and the surrounding contextual environments. A mature recommending decision should also be based on the contextual information of current learning activity, the user's profile, and the item's profile. Hence, as shown in Figure 3.4, the recommender system takes various types of information as input.

For other intelligent analysing modules (assistant plugins), they involved both user-side and item-side information (from the pre-processing module). The detail of the involved information might vary from task to task. For example, for the task of distraction-level prediction, the model needs to involve contextual information which could reflect a user's current surrounding environment, such as geolocation (learning in a library could have low distraction level than learning in a bus station). For the task of final grade prediction, the model needs to access and analyse a user's quizzes performance, the difficulty level of this course, and his/her historical learning behaviour of this course.

3.2.3 Isolated Problem for the Research Dataset

Considering the representative studies and experiments discussed in Chapter 2, except non-publicity, insufficient and inaccurate dataset problem for micro learning recommender system, isolated research dataset is another obstacle for the research of micro learning.

The datasets obtained from different sources are isolated. This has not been brought to sufficient attention from previous studies. Unlike research in some other domains, which often have standard datasets; the datasets from different online platforms are isolated, due to the vague information of learning resources and different curriculum structures. For research purpose, it is hard to use an auxiliary data source to supplement the target data source. For example, the study [Yang, et al., 2014] captures the textual information from the video content, and [Niemann, et al., 2013] uses the co-occurrence information to boost the collaborative filtering result. The textual information is useful in

mining semantic information among the learning resources, which may further boost the recommender system as proposed in the study [Niemann, et al., 2013]. However, because of the different sources, these two datasets cannot be fused directly. Although there are initiatives to push a non-profit sharing of research-oriented MOOC data [Lopez, et al., 2017]. Unfortunately, most data from several learning platforms (e.g., edX and Coursera) are still partially open to researchers, or merely open to their partners. Most research teams can but get access to very limited datasets. To solve the underlying awkward situation, researchers demand more complete and diverse data to drive the decision-making system.

3.3 Summary

In this chapter, the big picture of the proposed micro learning system is introduced and discussed. Firstly, we demonstrate the framework of the proposed system from the different perspective of views. Then, the data involved in the system is introduced from three aspects, which are the utilities of different types of input data, the data flow and their characteristics, and the problem of the research data in the micro learning area. Specifically, we discuss and analyse each intelligent module of the micro learning system, the inputs and the outputs of these modules, and how the information flow exchanges among these modules. We also give insight discussion of the different types of involved data.

Chapter 4 Pre-Processing

As mentioned in the previous chapters that a sophisticated micro learning system needs an intelligent pre-processing module to make the raw learning resources ready to be processed by other parts of the system. This intelligent module could consist of several A.I. based models for different pre-processing tasks. In my research, two pre-processing models are designed and evaluated, one for extracting key information and another for interpreting text content of the learning resources.

4.1 Information Extraction

In this research, a sequence labelling model is designed for automatically analysing the content of the information stream and then identifying, locating, and classifying the valuable information for the micro learning service. We herein propose a deep sequence labelling model (namely, deep Bi-LSTM-CNNs-CRF) for information extraction. This model tries to insightfully depict different aspects of the online micro learning scenario, summarises them together, and extracts valuable information for assisting the follow-up different intelligent processing modules of the micro learning service.

4.1.1 Model Design

In this section, we describe the design of the proposed deep sequence labelling model for information extraction. As mentioned in Chapter 2, it is an end-to-end model; its abstract-level workflow is shown in Figure 4.1. It takes the raw information as input and outputs the extracted key information. No intermediate output will be produced in this model. For the detail of the proposed model, firstly, from a high-level perspective, we introduce the overall architecture of this model. Next, for each

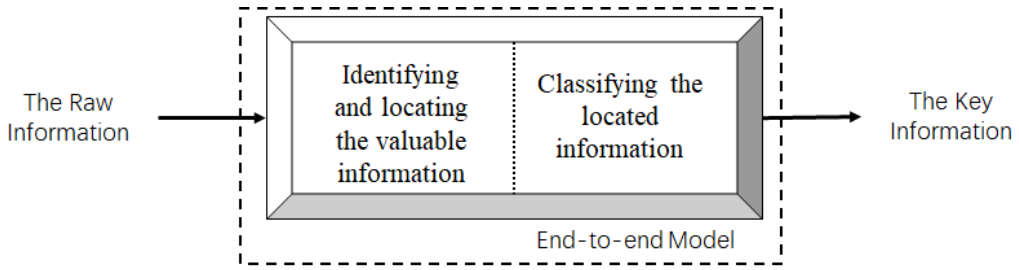

Figure 4.1 The Abstract-level of the End-to-end Information Extraction Workflow

low-level vital component of this model, we discuss its characteristics.

4.1.1.1 The Network Architecture

The proposed network contains four important layers and one block. The embedding layer is used for mapping the one- or multi-hot raw data to dense representations. The Bi-LSTM layer is used for modelling the temporal pattern of the embedded input sequence. The convolution neural network (CNN) layer is used for extracting adjacent latent features from the embedded input data. The CRF layer is used for adding extra local constraints, which are not captured by the previous layers. Moreover, the fusion block is used for combining different types of latent features. In the proposed network, the CNN layer and the Bi-LSTM layer model the embedded input separately. The input of this model is a sequence of vectors in the high dimension space. Each vector represents the selected features of an individual raw micro learning resource. The outputs of the CNN layer and the Bi-LSTM layer are jointly fed into the fusion block, which contains several non-linear transformation layers for better information fusion. The general architecture of the proposed neural network is shown in Figure 4.2.

4.1.1.2 Embedding Layer for Semantic Modelling and Dimension Reduction

The embedding technique is an effective method for dimensionality reduction and feature representation, which transfers the points lying in a high dimensional space to a low dimensional one while approximately preserving pairwise distances between points [Abdullah, et al., 2016]. Such technique shows satisfactory performance in reducing the data and model complexity, and has been used in various machine learning-related tasks, such as information retrieval [Mitra, et al., 2017], multimedia data processing [Wang, et al., 2018], and data mining [Zhou, et al., 2016]. For sequential

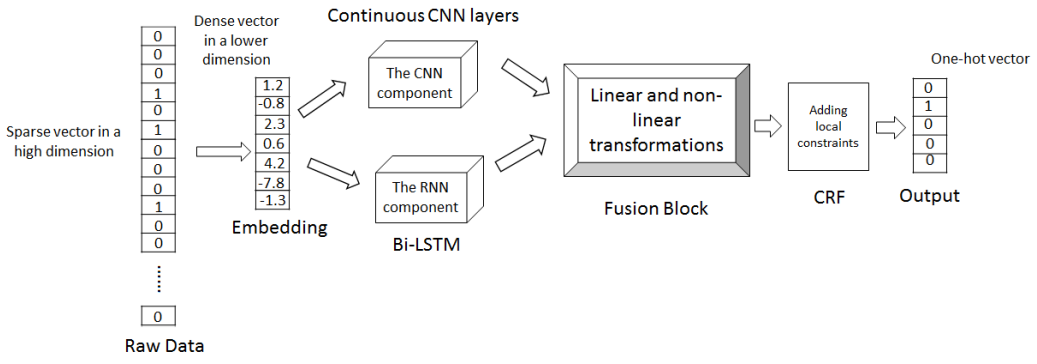

Figure 4.2 The Overall Network Structure of the Proposed Deep Bi-LSTM-CNNs-CRF Model

signal processing, especially in natural language processing (NLP), the embedding layer also contributes to modelling the semantic information of raw input data [Zhang, et al., 2015]. In our proposed model, an embedding layer is used to map sparse high-dimensional raw data into low-dimensional continuous dense one and extract the first-step semantic information.

4.1.1.3 CNN Layer for Latent Feature Extraction

Because of the competitive performance in extracting the latent feature, CNN has been widely used in the research area of computer vision and many other applications like NLP [Yin, et al., 2017]. Many previous studies such as [Ma, et al., 2016, Sutskever, et al., 2014, Zhai, et al., 2018], use CNN to extract and model character-level semantic information such as prefix and suffix.

However, not all types of information contain character-level semantic information. As a result, for constructing a more general model towards the information extraction task, after embedding layer, two continuous CNN layers are used to further mining and summarizing local features from adjacent inputs (object-level) in our proposed model. The utility of the CNN layers is quite different from the studies mentioned above. We may assume that such CNN layer can capture different types of information and boost the performance of Bi-LSTM-CRF. In computer vision, the sliding window strategy is widely used to restrict the amount of information involved in each summarization process [Papandreou, et al., 2015]; in our proposed model, we use a fix-size sliding window in each of the CNN layers. As shown in Figure 4.3, which is a single CNN layer for extracting latent features of adjacent embedded inputs, E_i is the i-th embedding vector generated from the previous embedding layer; the information from adjacent input embedding vectors is summarised. The rectangle is the sliding-window moving from left to right, and at each iteration, the information inside the window will be summarised.

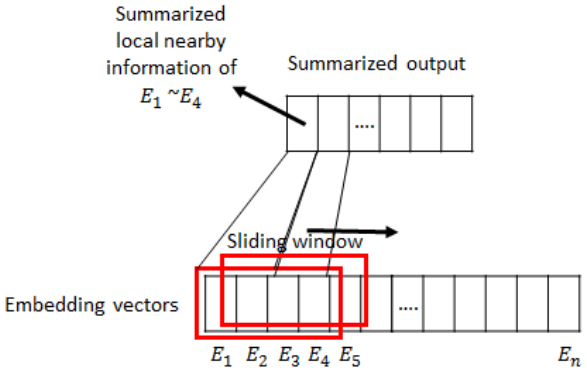

Figure 4.3 Convolutional Neural Network for Summarizing and Extracting Latent Features

4.1.1.4 Bi-LSTM Layer for Sequence Labelling

In our proposed model, one RNN layer is added as the core component after the embedding layer. This layer aims to extract and model temporal features. For a typical information extraction task, we may assume that both past and future inputs can provide valuable information in recognising and locating the information that needs to be extracted. Hence, in our model, a Bi-LSTM structure is used in the RNN layer. The workflow of Bi-LSTM is shown in Figure 4.4. The embedded information is fed into the Bi-LSTM layer in both the successive order and the reverse order; then for each time step, the Bi-LSTM layer will output a prediction based on the 'past' and 'future' information of the current 'moment'.

4.1.1.5 Fusion Block for Combining Different Types of Latent Feature

As discussed in [Chen, et al., 2017], properly fusing different types of features can make the information representation more reliable and more accurate. Inspired by this concept, a fusion block is added behind the CNN and Bi-LSTM components, and this block aims to better combine those different types of latent features. In the fusion block of the proposed model, latent features extracted from the CNN layer and the Bi-LSTM layer are firstly merged together by a concatenation operation. Then several non-linear transformation layers are used to further combine the latent features into fine-grained features. Such type of non-linear transformation layer could vary greatly respective to the problem domain and the complexity of the input data. In this study, we used another Bi-LSTM layer and a fully connected neural network to model this non-linear transformation. The structure detail of this fusion block is shown in Figure 4.5.

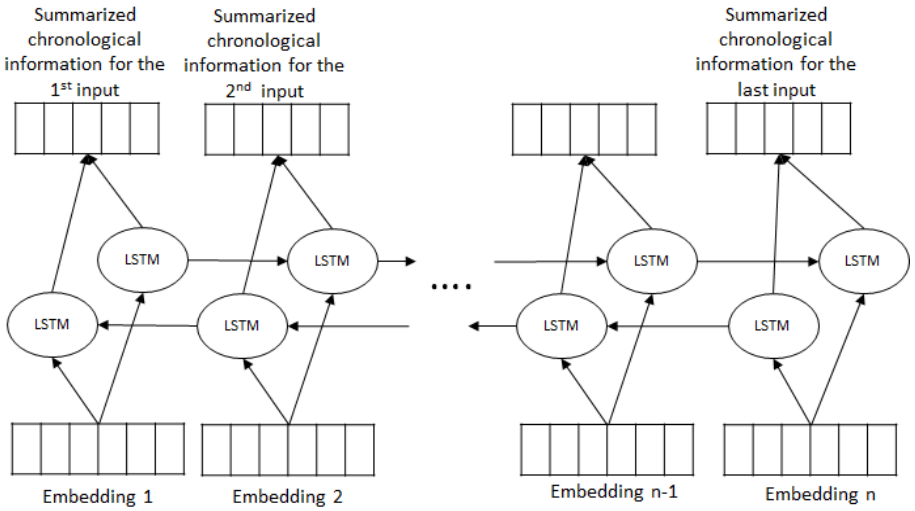

Figure 4.4 Bidirectional Long Short Memory for Modelling Sequential Pattern

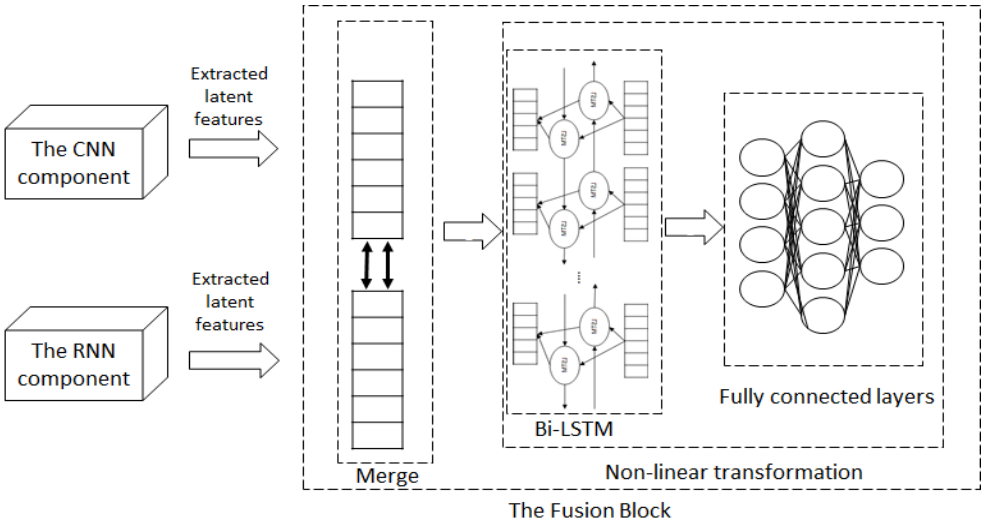

Figure 4.5 The Structure Detail of the Fusion Block

4.1.1.6 CRF Layer for Adding Local Constrains to the Sequential Model

As discussed earlier, a pure RNN model has its own disadvantage in modelling local constraints. Hence, a CRF layer is used prior to the final output layer of the whole model.

In the CRF model, for a given sequence x, the probability of output y could be simply formulated as Equation (4.1). From this equation, we can easily observe that y_i and y_{i-1} influences each other, indicating that this probability value considers the correlation between outputs in neighbourhoods. The network structure of CRF is shown in Figure 4.6 [Huang, et al., 2015]. Herein the prediction of the second output Y_2 is not only determined by the second input X_2 but also influenced by the first output Y_1. Function $\mathrm{t}(Y_{i-1}, Y_i)$ and $\mathrm{s}(Y_i, X_i)$ models the state transitions and emissions, respectively. This CRF layer aims to add more local constraints, especially the local constraints of the output sequence, which is not captured by former embedding, Bi-LSTM, and CNN layers [Ma, et al., 2016].

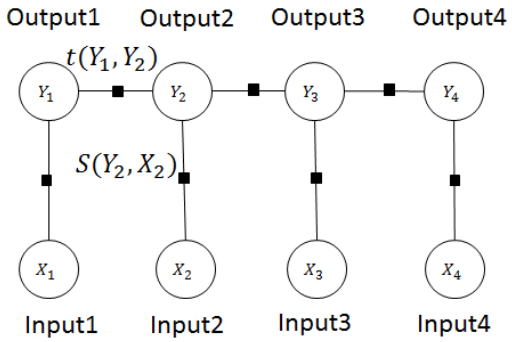

Figure 4.6 CRF Network

$$\mathrm{p}(\mathbf{y}|\mathbf{x}) = \frac{\prod_{i=1}^{n} \exp\,(y_{i-1}, y_i, \boldsymbol{x})}{\sum_{y \in y(x)} \prod_{i=1}^{n} \exp\,(y_{i-1}, y_i \boldsymbol{x})} \tag{4.1}$$

4.1.2 Experiment and Analysis

We compare the performance of CRF, LSTM, Bi-LSTM, and Bi-LSTM-CRF with our proposed deep Bi-LSTM-CNNs-CRF model using different experiment settings. We use three different evaluation metrics: precision, recall, f1-score, and the area under curve (AUC) value.

4.1.2.1 Evaluation Metrics

Precision is the number of true positive predictions divided by the total number of positive elements that predicted, which reflects how accurate the model is getting out of the predicted positives. The recall is the number of true positive predictions divided by the total number of positive elements in the data set, reflecting how many of the actual positives that the model predicts through all the positives. F1-score considers both the precision value and the recall value, which can better reflect the model performance when the class distribution is uneven. The receiver operating characteristic curve (ROC) is about plotting the true positive rate (TPR) against the false positive rate (FPR). The area under curve (AUC) can reflect the ability of how much a model can distinguish different information.

4.1.2.2 Dataset

The data used in the experiment comes from different fields of documents where we aim to facilitate the information extraction task of an online learning service that covers a variety of disciplines. There are two separate datasets used in this study; one is a labelled dataset for training the model except for the embedding layer, which contains four different types of information that need to be extracted. Another dataset is unlabelled, which is used for training the embedding layer, whereas the training is unsupervised. The unlabelled dataset is about a hundred times bigger than the labelled counterparts. Both datasets are encoded sequential of data and coming from the same source.

For a single case of a certain online learner, most of the online information is redundant. As a result, in this experiment, the information that needs to be extracted is very sparse. Also, in order to better reflect the real-world task where information with different significant levels or in different types could have totally different distributions, in the datasets, these four types of information are not evenly distributed. The statistical information about each type of information involved in the dataset

is shown in Table 4.1.

Moreover, to maintain the generalization ability of our proposed deep Bi-LSTM-CNNs-CRF model, no handcraft domain-relevant feature is involved in this study. All the features that contain domain information are automatically selected, or extracted, or summarized by different layers of the neural network. There is no pre-trained model used in this study, even some of them are powerful and can greatly simplify the training process, such as Bert [Devlin, et al., 2019] and ELMo [Peters, et al., 2018]. Our goal is to demonstrate that the proposed model is not restricted to any domain-relevant datasets or any domains. And how to effectively apply it to a specific online educational scenario or a problem domain is beyond the scope of this paper and will be considered as future work.

4.1.2.3 Experimental Setup

In order to make these models comparable, we fixed the hyper-parameters of each neural network component during the training process of each model based on the pre-experiment results. The embedding technique used in this study is the classic unsupervised word2vec model [Mikolov, et al., 2013], and the embedding dimension is 32. The output dimension of the first CNN layer is 128 and the second CNN layer is 64, and the sliding window size for both CNN layers is 5. The output dimensions of the Bi-LSTM outside and inside the fusion block is 128 and 64, respectively, and the maximum sequence length for modelling the chronological pattern is 128. All the neural network layers held a 0.3 dropout rate, and we used default initialized settings for other parameters such as activation function and weight initialization method. Ten-fold cross-validation is used in the experiments.

4.1.2.4 Experiment Results and Discussions

The experiment results of the proposed model for information extraction are shown in Table 4.2. To

Table 4.1 Statistical Information about the Dataset

	Information A	Information B	Information C	Unwanted Information	Total
Number	45,736	34,655	26,221	761,237	867,849
Percentage	5.26%	3.99%	3.02%	87.71%	100%

better represent the performance of each model, despite the recall, precision, and f1-score, we also involved the total number of information that each model extracted containing true positive (TP) and true negative (TN) results. The total number of ground truth information that needs to be extracted is 7,345.

4.1.2.4.1 The Significance of CRF Layer

Although the neural network has demonstrated its superb ability in modelling complex problems, based on our experiment results, we can easily find out that the pure CRF model greatly outperforms pure RNN-based models (LSTM and Bi-LSTM). This finding highlights the significance of the CRF layer, especially for the problems where the adjacent outputs have strong correlation or connections with each other.

In this study, we also compare the performance of the CRF method and the widely used softmax function. We keep all the settings of the model constant, only replacing the final CRF layer with the softmax function. From the bottom two rows of Table 4.2, we can observe that, for extracting information from the information sequence, the CRF layer is greatly outperforming the softmax function.

4.1.2.4.2 The Bi-directional RNN and the CNN Layer

The experiment results show that the pure Bi-LSTM model performs better than the pure LSTM model. This result confirms our assumption in Section 4.1.1. For a general sequential modelling problem, ‘future information’ does help the model make more accurate predictions and, therefore, should be involved in the model.

From the last column of Table 4.2, we can find that the number of information extracted by Bi-LSTM is higher than CRF and LSTM, but the performance of Bi-LSTM is much worse than CRF. This indicates that Bi-LSTM can extract more information than LSTM, but it is also prone to involving more true-negative information. In general, Bi-LSTM has the potential to extract more diverse information but requires a constraining module or layer to filter out the true-negative predictions.

The model combining Bi-LSTM and CRF shows a huge leap in the recall, precision, and F1 score. This proves that the Bi-LSTM-CRF model can maintain the strength of the pure CRF model and

Table 4.2 Experiment Results of Different Models

Metrics / Model	Recall	Precision	F1	# Results (TP+TN)
CRF	0.5164	0.3682	0.4299	5,238
LSTM	0.3016	0.1994	0.2401	4,858
BiLSTM	0.3231	0.2509	0.2824	5,704
BiLSTM-CRF	0.8168	0.7908	0.8089	7,113
Proposed model (without fusion block)	0.8083	0.7851	0.7951	7,136
Proposed model	**0.8223**	**0.8188**	**0.8206**	**7,316**
Proposed model (use softmax function to replace the CRF layer)	0.8052	0.7897	0.7974	7,205

pure Bi-LSTM model and eliminate the drawbacks of these two models.

The model proposed in this study outperforms the Bi-LSTM-CRF model. This result suggests that adding a convolution layer and then using appropriate fusion block is really achieving the extraction of some latent information, which is not captured by former embedding, RNN and later CRF layers. We maintain the reason for the improvement of the model is towing to the structure difference of the CNN layer and the RNN layer, and this makes it easier to model and represent different aspects of the same problem.

4.1.2.4.3 Viterbi Algorithm and Cross-entropy

This study also compares the model performance between using the Viterbi algorithm combined with negative log-likelihood and using cross-entropy for calculating the loss during the model optimisation stage. Viterbi algorithm is mainly used for dynamically searching the best path for sequential predictions.

The performances of three representative models, pure CRF, Bi-LSTM-CRF, and the proposed model with different optimisation strategies are shown in Table 4.3. The results demonstrate that using the Viterbi algorithm combined with negative log-likelihood for optimising the model outperforms using cross-entropy. Simply using cross-entropy is prone to involving more true-negative predictions, which also indicates that, the Viterbi algorithm adds some constraints to eliminate wrong predictions during the prediction process, as well.

Table 4.3 Comparison of the Viterbi Algorithm and Cross-entropy

Metrics / Model	Recall	Precision	F1	# Results (TP+TN)
CRF (Viterbi)	0.5164	0.3682	0.4299	5,238
CRF (Cross-entropy)	0.2535	0.1999	0.2236	5,795
Bi-LSTM-CRF(Viterbi)	0.8168	0.7908	0.8089	7,113
Bi-LSTM-CRF(Cross-entropy)	0.7367	0.7319	0.7343	7,299
Proposed model (Viterbi)	0.8223	0.8188	0.8206	7,316
Proposed model (Cross-entropy)	0.8057	0.8056	0.8057	7,346

This phenomenon is because the outputs are not independent, and the Viterbi algorithm has a better ability to find the best path for the sequential output.

4.1.2.4.4 The Significance of the Fusion Block

Moreover, by comparing the model with the fusion block and without the fusion block, we can see that the model with the CNN layer performs even worse than the Bi-LSTM-CRF model without the fusion block. The model extracts more true-negative information when adding the CNN layer but without using fusion block. This result indicates that, even though the CNN layer can extract some other latent information, the model still requires a proper information fusion procedure to combine different aspects of information. Simply adding a fusion block does increase the recall, precision, and F1-score in this experiment.

4.1.2.4.5 The Information Distinguishing Ability

As discussed in the previous section, the distribution of different types of information could vary greatly. For a personalized online learning service like micro learning, the ability to distinguish and extracting different types of information is significant for capturing characteristics of both learners and learning materials. From the other perspective to demonstrate the robustness of the proposed deep Bi-LSTM-CNNs-CRF model, we also compared such ability of representative models on different types of information. The overall information distinguishing ability of these models is

shown in Figure 4.7.

The proposed deep sequence labelling model greatly outperforms the pure CRF model and pure Bi-LSTM model, and the AUC score of our proposed model is about 2% higher than the mainstream Bi-LSTM-CRF model. For different types of information, the proposed model shows satisfactory performance than any other models. This result indicates that our model has great robustness, which can precisely locate and classify different types of information with different distributions. The details of the model performance in distinguishing our pre-defined types of information are shown in Figure. 4.8, where the class 1 to 3 represent the information A, B, and C, respectively (Table 4.1); the class 0 represents the useless information. The distributions of these types of information are shown in Table 4.1.

4.1.2.4.6 The Efficiency of the Proposed Model

Lastly, in this study, we also compare the training efficiency of the three most complex models, Bi-LSTM-CRF, the proposed model, and the proposed model without fusion block. Due to the simplicity of the model structure and less satisfying performance, the training efficiency of pure CRF,

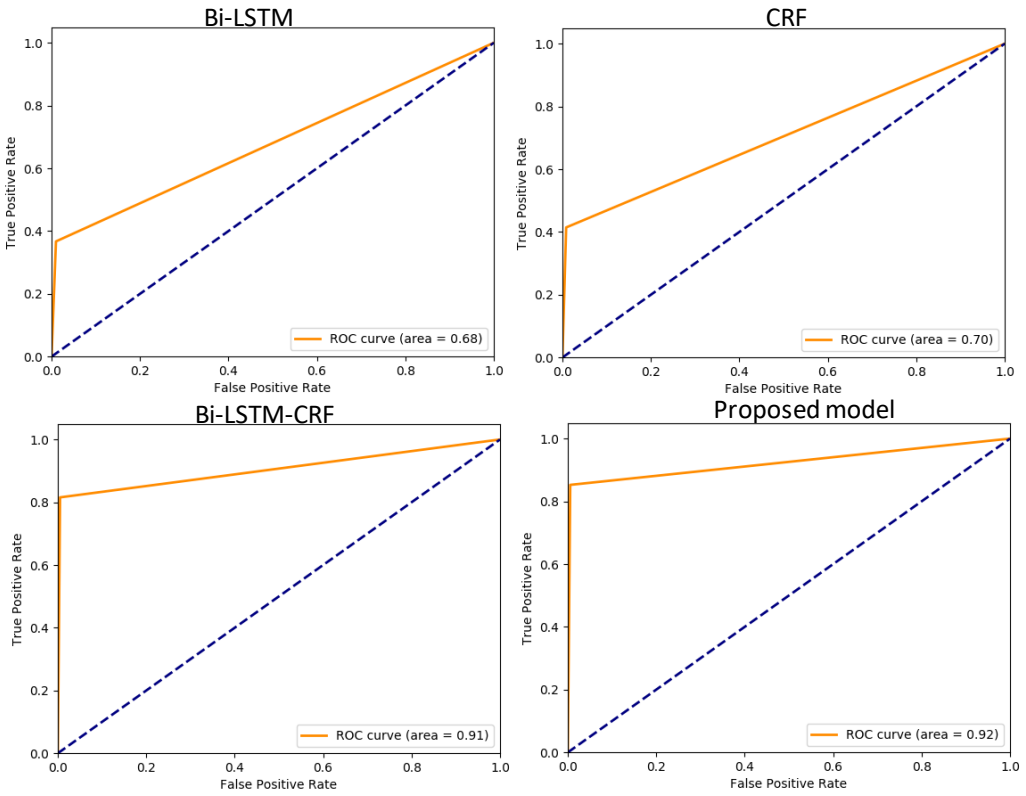

Figure 4.7 The AUC of Total Extracted Information

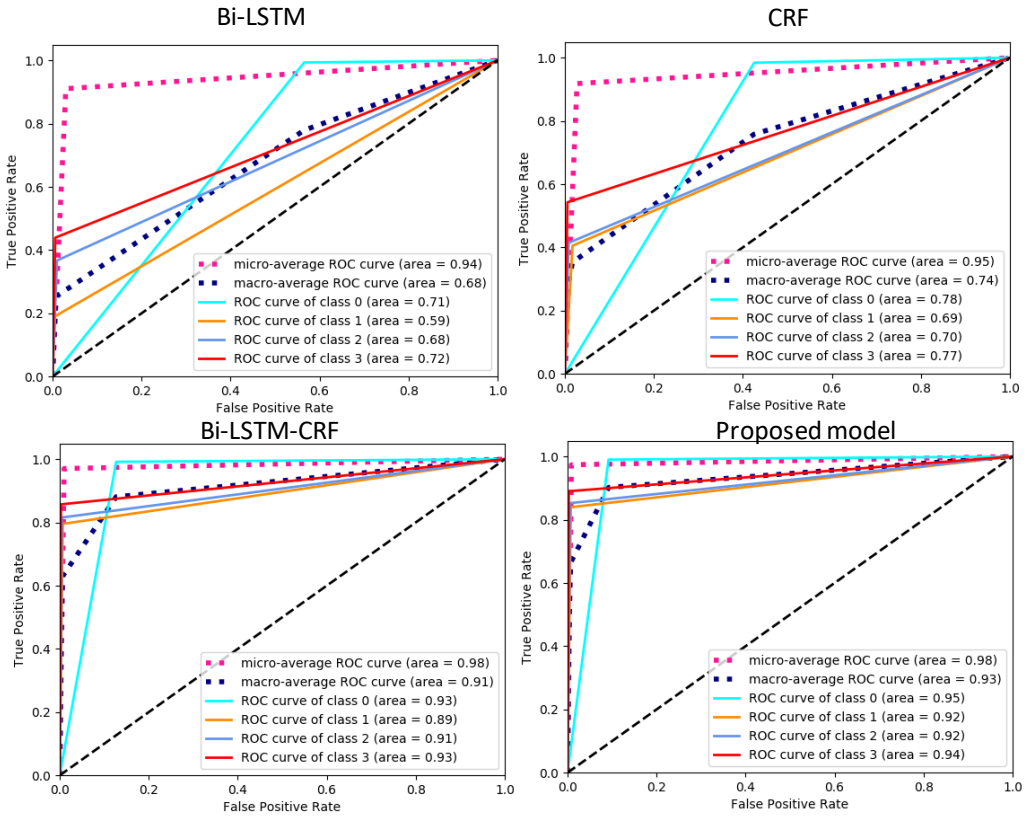

Figure 4.8 The ROC and AUC of Different Types of Extracted Information

LSTM, Bi-LSTM are not compared in the experiments. The decrease of the loss value during the training process is shown in Figure 4.9. According to Figure 4.9, three models converge within similar training steps. Comparing with the mainstream sequence labelling model Bi-LSTM-CRF, the proposed model with additional CNN layer and fusion block does not show any obvious drop in the

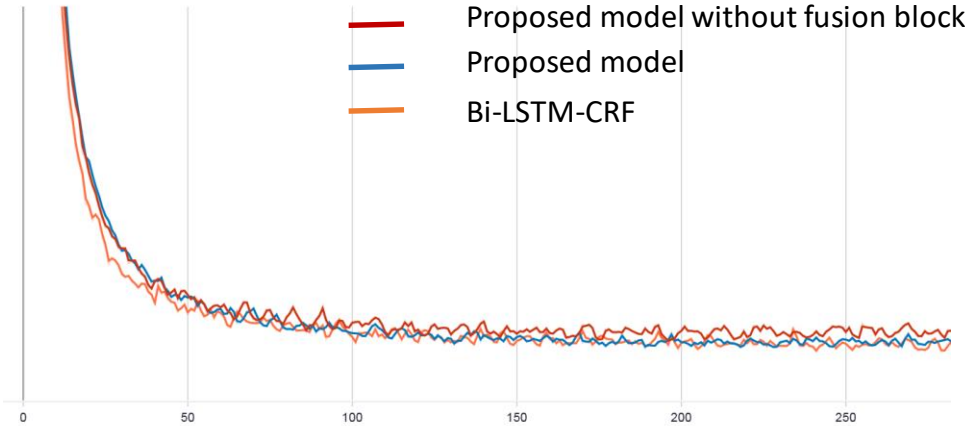

Figure 4.9 The Changes of the Loss Values for Each Training Epoch.

(Vertical and horizontal coordinate refers to the learning loss and learning epoch, respectively)

training efficiency. And the difference in training efficiency between the proposed model with and without fusion block is also very unnoticeable. Hence, our model does not require more extra training steps to reach an optimal state.

4.2 Text Analysis

The online learning resource contains various types of short text contents, such as course title, course description, and learners' comments. These text contents can provide useful information about the resources themselves and user experience to different decision-making modules of a micro learning system. Compared to the well-structured formal text content (such as course description), the comments given by the online learners are hard to interpret, as they are informal and contain much rich information. Hence, in this study, a deep learning model is designed to interpreting the content of the informal short text. This part of the research aims to develop a universal deep learning model which has the potential to be applied to other formats of information (like images) with little technical adjustments.

4.2.1 Model Design

In this section, we describe the design of the proposed CNN-based framework for interpreting the content of the short text. Firstly, from a high-level perspective, we introduce the overall architecture of this framework. Next, how the research problem is formulated in this study will be demonstrated.

4.2.1.1 The Architecture Framework

The proposed framework contains two components which are the upstream component and the downstream component. The general architecture of the proposed framework is shown in Figure 4.10. The upstream component is a pre-trained language model. It is used to transform the raw text input into dense embeddings. The downstream component is a task-specific model, which takes in the dense embeddings and produces the final predictions for a specific NLP task. In this study, we

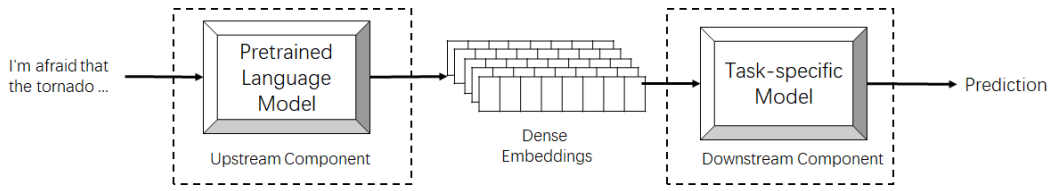

Figure 4.10The Framework Architecture

only focus on applying CV ideas to the downstream component, whereas designing the upstream component is beyond the scope of this research as many studies have been carried out in this field. However, to get comprehensive experiment results, different pre-trained language models will be compared and discussed in this chapter later.

4.2.1.2 Problem Formulation

To clearly formulate the task of informal short text content understanding, we would have the following definitions.

4.2.1.2.1 Upstream Component

In this study, the pre-trained language model L is used to generate dense word embeddings $\boldsymbol{e}$ with dimension *m*. For each word *w* from a sequence of text $\boldsymbol{t} = (\mathrm{w}_1, \mathrm{w}_2, \ldots \mathrm{w}_\mathrm{i})$, the mapping procedure can be formulated as $\boldsymbol{L}(\mathrm{w}_\mathrm{i}) => \mathrm{e}_\mathrm{i}, \mathrm{e}_\mathrm{i} \in \mathbb{R}^\mathrm{m}$.

4.2.1.2.2 Text Representation

In this study, a sequence of text is represented in the form of 2D 'figure' P by stacking all embeddings together, such procedure is formulated as $\boldsymbol{P} = Stack(e_1, e_2, \ldots e_i)$. To prevent any information loss in this procedure, the order of the embeddings is kept the same with the original text sequence. The illustration of this process is shown in Figure 4.11.

4.2.1.2.3 Downstream Component

Given different NLP tasks have different goals, the designs of the downstream component can vary from task to task. The downstream component takes in generated text 'figure' $\boldsymbol{P}$ and makes the final

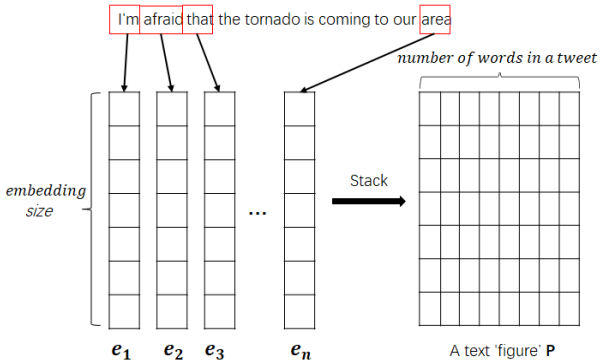

Figure 4.11 The Organization of a Piece of Text 'Figure'

prediction y. Two different datasets are used in this study, one for topic prediction and another for sentiment analysis. For a topic prediction task, the goal is to analyse the content of a given sequence of text $\boldsymbol{t}$ and predict whether this sequence of text is about the target topic or not. For this task, defining the ground-truth as $y_{\text{topic}} \in \{0, 1\}$, where 1 is for the target topic and 0 is for not about the target topic. For the sentiment analysis task, the goal is to find out whether a given sequence of text $\boldsymbol{t}$ contains a certain sentiment or not, and what its sentiment is. For this task, defining the ground-truth as $y_{\text{sentiment}} \in \{0, 1, 2\}$, where 0 is for neutral (no obvious sentiment involved in a sequence of text), 1 is for negative sentiment, and 2 is for positive sentiment. Hence, together with the above two definitions, the goal of the downstream component is to learn the following Equation (4.2) $\mathcal{F}$:

$$\mathcal{F}(\mathbf{p}) \Longrightarrow y \tag{4.2}$$

Inspired by the prior work [Kim, 2014] and the idea of the n-gram model [Roark, et al., 2007], we have carefully designed a CNN-based network to interpret the text content. The generated '2D' text figure is scanned by multiple kernels $k_1 \sim k_i$ serval times to extract semantic information. Different kernels have different widths d_i but share the same height h. The kernel height h is equal to the word embedding size. With these settings, kernels can summarise the information of number of d_i successive words at each step during the first convolutional operation, such procedure is similar to generating n-gram samples. The following convolutional operations extract different levels of granularity of information in the same way as it has been frequently used in the CV area. After a set of successive convolutional operations, pooling operation and fully connected layers are applied to summarise all extracted information and produced the final predictions. The illustration of the network structure of the proposed CNN-based downstream component is shown in Figure 4.12.

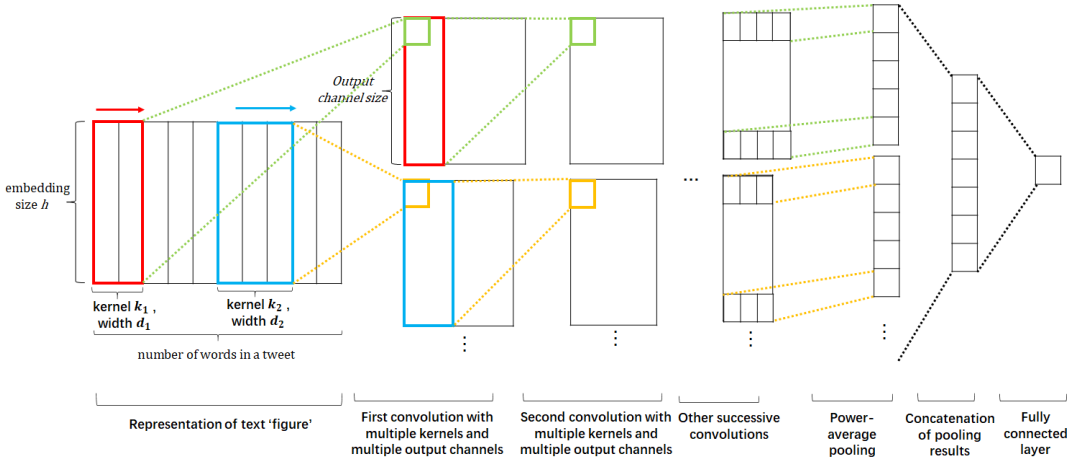

Figure 4.12 Network Structure of CNN-based Downstream Component

4.2.1.2.4 1D-Bounding Box

Moreover, in this study, despite making the proposed model understand informal short text content, we also investigate whether the proposed model can identify which phrases or words in a sequence of text would express the key information. In another word, for the sentiment analysis task, we want the model to be able to identify which words or phrases could express a negative or positive sentiment for a given sequence of text. Inspired by the utility of the bounding box in object detection, a 1D 'bounding box'[5] is used to mark the key information. The example of the bounding box in both areas is shown in Figure 4.13. In Figure 4.13, the left part is the bounding box of the object detection results of cats; the right part is the bounding box of the sentiment analysis result of negative feeling. The 1D bounding box is formulated as $\boldsymbol{B}(\mathrm{c}_{\mathrm{index}}, \mathrm{l})$, the first element is the center index of the selected sequence of text and the second element is the length of selected text. The loss function to measure the prediction and the ground-truth of the bounding box is formulated as Equation (4.3):

$$Loss_B = MSE(\mathrm{c}_{\mathrm{index}} - \widehat{\mathrm{c}_{\mathrm{index}}}) + MSE(\mathrm{l} - \hat{\mathrm{l}}) \tag{4.3}$$

where the MSE is the mean square error loss. Difference between the predicted centre and the ground-truth centre is measured by the first term, and the difference between the predicted length of key words/phrase and the ground-truth length is measured by the second term.

4.2.2 Experiment and Analysis

The experimental details are demonstrated and analysed in this section, including the introduction of the datasets, the used evaluation metrics, the baselines, the settings of the experiment and the analysis of the results.

4.2.2.1 Datasets

As there is no public informal short text from the educational domain for the research usage, the

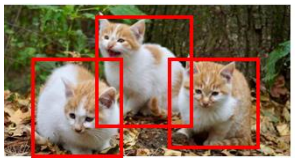

Figure 4.13 The Example Representation of Bounding Box in CV (left) and NLP (right)

[5] The working manner and optimization process of the proposed bounding box is different from the one in computer vision. We merely use a similar idea to identify the wanted information.

datasets used in this study are collected from the social media platform. Two short text datasets are used in the experiments; and both are collected from the Twitter platform and open to the public[6]. The first dataset contains more than 10 thousand tweets, some of which talk about real disaster events. The ratio of disaster-related against non-disaster-related is 43:57. This dataset contains the raw text of each tweet, keyword from the tweets, and location information where a tweet was sent from. The keyword and location information may be blank; we only utilise raw text information to train the model. The second dataset contains 30 thousand tweets with or without sentiment information. The ratio of the number of neutral against positive against negative sample is 41:31:28. This dataset contains the raw text of each tweet and the text fragment that support the tweet's sentiment. Similarly, only the raw text is used to train the model. The raw text is collected directly from the Twitter platform and has not been pre-processed yet. The example of these two datasets is shown in Table 4.4.

4.2.2.2 Evaluation Metrics

In order to reflect the model performance from different perspectives, three different types of evaluation metrics are used in the experiment. The first evaluation metric is accuracy (Acc), formulated as Equation (4.4), which directly reflects the proportion of the correct predictions produced by each model. However, for imbalanced distributed ground-truth, this metric might not be suitable for comparing the effectiveness of the models [Valverde-Albacete, et al., 2014].

$$Acc = \frac{total\ number\ of\ correct\ prediction}{total\ number\ of\ prediction} \tag{4.4}$$

The second evaluation metric used in this study is the Area Under Curve (AUC) value. AUC value reflects the ability of a model to distinguish different types of information (i.e., for task one about whether it is a disaster or not, for task two about whether they are different sentiments). When dealing with the multi-class classification task (the task of sentiment analysis), the one-versus-one strategy was used in the experiment. AUC value is the area under the Receiver Operating Characteristic (ROC).

[6] The dataset about disaster prediction was created by the company figure-eight and originally shared on https://www.figure-eight.com/data-for-everyone/. The second dataset about sentiment analysis was extracted from https://appen.com/resources/datasets/.

Table 4.4 Data Sample of the Two Dataset

Dataset	Information		
Disaster Prediction	Raw text	Keyword	Location
	I-77 Mile Marker 31 to 40 South Mooresville Iredell Vehicle Accident Congestion at 8/6 1:18 PM	*accident*	*North Carolina*
Sentiment Analysis	Raw text	Text fragment	
	A little happy for the wine jeje ok it`sm my free time so who cares, jaja i love this day	*A little happy*	

The last evaluation metric is the F-score, formulated as Equation (4.5), which is the harmonic mean of the Recall score, formulated as Equation (4.6), and the Precision score, formulated as Equation (4.7). Herein, *TP* is the number of true-positive results, *FN* is the number of false-negative results and *FP* is the number of false-positive results. Because of the trade-off between Recall and Precision, we cannot conclude a good model merely based on a high Recall score or high Precision score. Hence, using F-score is a better choice for model comparison.

$$F_1 = 2 \times \frac{recall \times precision}{recall + precision} \tag{4.5}$$

$$recall = \frac{TP}{TP+FN} \tag{4.6}$$

$$precision = \frac{TP}{TP+FP} \tag{4.7}$$

4.2.2.3 Baselines

In the experiments, different pre-trained models are used to investigate the effectiveness of applying the CV solution to the NLP problem. Specifically, for the upstream component, we involve the following pre-trained language models:

1. Word2Vec [Mikolov, et al., 2013]: This model has high optimization efficiency, but can only model the local semantic information within the pre-defined window.
2. GloVe [Pennington, et al., 2014]: This model combines the merits of LSA [Deerwester, et al., 1990] and Word2vec. It uses the co-occurrence matrix to model the local and global semantic information at the same time.
3. Bert [Devlin, et al., 2018]: Bert and its variants dynamically model the semantic information. As indicated in the original study that a multitask fine-tuning approach could be used to train

the model, and this would boost the performance even further.

For the downstream component, we compare the following models' effectiveness in understanding tweet content:

1. Bi-GRU: Bi-directional gated recurrent unit neural network.
2. Bi-LSTM: Bi-directional long-short-term-memory neural network.
3. The proposed CNN-based model in this research.

4.2.2.4 Experimental Setup

In this research, all the models are implemented using PyTorch framework [Paszke, et al., 2019]. The pre-trained language models Bert and Word2Vec are implemented through Transformer[7] and Gensim[8], respectively; the pre-trained GloVe model is reproduced through its pre-trained word vectors[9]. Six different kernels with 128 output channels are used in the proposed CNN-based model. ReLU is used as the activation function for each convolutional output, and there are four successive convolutional layers in total. The dimension number for the hidden layer of Bi-LSTM and Bi-GRU is set to 256. The sigmoid function is used to produce the final prediction for the disaster prediction task, while the softmax function is used to produce the final prediction for the sentiment analysis task. All the other settings strictly stick to the original work, or we directly use the default settings of the PyTorch framework. The early-stop mechanism is applied to all the training processes to prevent overfitting.

Before using the language model to convert the tweet content to dense vectors, the raw text is pre-processed through several NLP data cleaning and normalizing stages, ranging from removing the stop-words, lemmatization, to removing URLs and emojis.

4.2.2.5 Experiment Results and Discussions

Table 4.5 illustrates the effectiveness comparison of different downstream components for two different NLP tasks. Table 4.6 reports the effectiveness of different upstream components. As we have obtained similar results from two datasets, we only present the experimental results of the task disaster prediction (topic detection) in Table 4.6. The demonstration of the 1D bounding box to locate key information is shown in Table 4.7.

[7] https://huggingface.co/transformers/index.html
[8] https://radimrehurek.com/gensim/#
[9] https://nlp.stanford.edu/projects/glove/

4.2.2.5.1 The Effectiveness of the Proposed CNN Model

According to the results in Table 4.5, we can easily conclude that, with the same Bert upstream component, the proposed CNN-based downstream component shows competitiveness comparing to mainstream NLP solutions in all criteria for two different tasks (highlighted in bold text). For the task of disaster prediction, the Bi-LSTM model slightly outperforms the Bi-GRU model, while for the task of sentiment analysis, the Bi-GRU model greatly outperforms the Bi-LSTM model. We would argue such improvement was produced by the structure difference between the LSTM cell and GRU cell. According to the original work of LSTM [Gers, et al., 1999] and GRU [Cho, et al., 2014], the LSTM tends to remember longer semantic information than GRU. Hence, we consider that for different tasks, involving too much information during the modelling procedure will not always improve the model performance.

The task of disaster prediction needs to understand the whole tweet to predict whether the given tweet is about disaster or not. It is hard to infer whether a tweet is about a disaster or not merely based on a short text segment or phrase. As shown in the following example:

"All residents asked to 'shelter in place' are being notified by officers. No other evacuation or shelter in place orders are expected"

Remembering more words or longer text sequence is helpful to understand the context of the whole tweets. Hence, for such a task, the LSTM-based model is better than the GRU-based model.

As for the task of the sentiment analysis, in most cases, sentiment is just one part of tweet content.

Table 4.5 Downstream Component Comparison on the Tasks of Sentiment Analysis and Disaster Prediction

Disaster Prediction			
Models	Acc	F1	AUC
Bert + Proposed	**0.8238**	**0.8169**	**0.8798**
Bert + Bi-GRU	0.8108	0.8044	0.8698
Bert + Bi-LSTM	0.8172	0.8128	0.8695
Sentiment Analysis			
Models	Acc	F1	AUC
Bert + Proposed	**0.8573**	**0.8354**	**0.8998**
Bert + Bi-GRU	0.8465	0.8273	0.8800
Bert + Bi-LSTM	0.8159	0.7927	0.8543

As a negative sentiment tweet shown in the following example, a user posts a certain event and expresses how he/she feels about it:

"Grrr..stupid internet connection ruined a great scrabble game"

Interpreting the sentiment of a user relies on a short text sequence or phrase. Inferring from the longer text sequence might negatively affect the model performance, as the phrase "a great scrabble game" shows positive sentiment in the above example, which contains negative sentiment. Hence, with the same settings, the Bi-GRU model greatly outperforms the Bi-LSTM model in the second task.

The above long-short-sequence modelling problem can be avoided by using the proposed CNN-based model. Using the CNN-based model, we can flexibly control the length of the word sequence to be modelled at each step by setting the kernel size at the first convolutional layer. The configuration of the kernel sizes can be determined on domain knowledge or pilot experiments.

4.2.2.5.2 The Importance of Upstream Language Model

We also investigate the framework performances in using different language models in the upstream component. From Table 4.6, we can conclude that the frameworks with the pre-trained Bert model outperform the ones with the GloVe model or Word2Vec model (highlighted in bold text). As mentioned in Section 4.2.2.3, the Bert model can capture more semantic information. A better language model indicates the downstream component can access more useful information. From the

Table 4.6 Upstream Component Comparison on the Task of Disaster Prediction

Models	Acc	F1	AUC
Bert + Proposed	**0.8238**	**0.8169**	**0.8798**
GloVe + Proposed	o.8173	0.8145	0.8723
Word2Vec + Proposed	0.6852	0.6762	0.7271
Bert + Bi-GRU	**0.8108**	**0.8044**	**0.8698**
GloVe + Bi-GRU	0.7667	0.7598	0.8097
Word2Vec + Bi-GRU	0.6819	0.6760	0.7348
Bert + Bi-LSTM	**0.8172**	**0.8128**	**0.8695**
GloVe + Bi-LSTM	0.8107	0.8039	0.8588
Word2Vec + Bi-LSTM	0.6770	0.6625	0.7283

CV perspective, a better language model can generate 'higher resolution text figure', which is critical for mining details from the 'text figure'. As argued in one prior study [Neubig, et al., 2013], tweets generally contain more information per character, likely a result of Twitter-specific abbreviations and a less consistent writing style. Hence, to better interpret the tweet content, it is necessary to use a powerful language model for maximizing the retention of semantic information.

4.2.2.5.3 Locating Key Information

Making use of the second dataset, we also further investigate whether our proposed model can identify and locate the key information (i.e., keywords). As mentioned in section III, a 1D-bounding box is designed to select the keywords (for an NLP task, such job is usually accomplished by using the attention mechanism). Due to the space limitation, in this thesis, we only randomly present five positive and five negative results as shown in Table 4.7. The column of '**Ground-truth Key Words**'

Table 4.7 Demonstration of 1D Bounding Box on the Task of Sentiment Analysis

Raw Text	Selected Text	Ground-truth Key Words	Sentiment
Grrr..stupid internet connection ruined a great scrabble game	grrr..stupid	Grrr..stupid internet connection	negative
listening to the best days of your life by kellie pickler	listening to the best days	listening to the best days of your life	positive
awesome! All deserved I`m sure. Miss the Crabs games	awesome	awesome	positive
Are you going to hate being around my baby?	are you going to hate being	hate	negative
awww I love me some charlies we are enjoying some lucky food LOL	awww i love me some	love	positive
NICE! Got any that are indexed that you want to unload? I need a few.	nice! got any that	NICE!	positive
Still gutted that man utd lost	Still gutted that man utd lost	Still gutted that man utd lost	negative
i miss him ALOT but im not gonna talk to him, i HOPE	i miss him alot but im not gon na talk to him	miss	negative
Starting to spoil my pug since her brother Max passed away on Tuesday. We miss him.	starting to spoil my pug	We miss him.	negative
Just try to do your best. I hope you don`t get laid off.	try to do your best.	Just try to do your best. I hope you don`t get laid off.	positive

shows the labelled ground-truth words containing sentiment information. The column of 'Selected Text' is showing the keywords selected by the bounding box. We can clearly see that our proposed model can not only understand the sentiment meaning of the tweet content but also identify which words express such meaning. The only drawback of our bounding box is that it is prone to selecting longer text than the ground truth (by comparing the second and the third column). We attribute it to the common situations that longer text could contain more information and would be more supportive of the bounding box to make a decision.

4.3 Summary

In this Chapter, two novel models are proposed for pre-processing the raw information of the massive online learning resources.

4.3.1 Conclusion of Information Extraction and the Future Direction

Firstly, we propose a deep Bi-LSTM-CNNs-CRF model to extract valuable information from massive and redundant information streams for micro learning services. We compare our proposed model with several classical and widely adopted sequence labelling models. The experiments conducted in this study demonstrate the robustness of the proposed model. The model can identify new latent features and outperforming the mainstream Bi-LSTM-CRF based model.

From our experimental results, we can conclude five specific points about information extraction or sequence labelling: 1. the CRF layer is vital for sequence modelling; for some cases, pure CRF model performs even better than pure RNN models such as LSTM or Bi-LSTM; 2. for a general application case, using bi-directional RNN such as Bi-LSTM is a better choice than using single direction RNN; as 'future' input can anyhow provide some information for the sequence modelling; 3. CNN is useful in mining supplementary information and further boosting the performance of the current model; 4. the model proposed in this study demonstrates that it has the ability to efficiently mine useful information for online service. 5. compared with other representative models, our model shows satisfactory characteristics in both efficiency and robustness.

As discussed in a recent study [Wu, et al., 2015], a learning service always has underlying pedagogical issues. Different subjects or disciplines present different contexts and may require significantly different information. Hence, in order to better adapt to a general micro learning service,

it is pertinent to involve more global information in the information extraction process. In the future, it is necessary to continue improving the proposed information extraction model by mining and involving global information about the education and learning domain.

4.3.2 Conclusion of Text Content Analysis and the Future Direction

In this chapter, we investigate the feasibility of designing a generic model to solve the multimodal information problem. A CNN-based model and a further bounding box are proposed to demonstrate that, with proper adjustments, the mainstream solutions from the CV area can also be used to solve NLP problems with different goals. With proper configurations, the proposed CNN-based model can yield a better generalization ability than the RNN-based model. Based on the experiment results, we discover the significance of the language model in modelling textual information. The experimental results have also shown that our proposed solution has competitive performance compared to mainstream NLP solutions such as Bi-LSTM and Bi-GRU for the task of short text understanding. Moreover, the proposed model shows satisfactory performance in locating key information.

In the future, firstly, we will continue investigating the effectiveness of the proposed solution for long text understanding. To better design such a generic model for processing multimodal information, we will also investigate in applying CNN-based solution to solve other forms of information such as audio signals. In the meantime, we will also try to use NLP solutions reversely to solve other forms of information.

Chapter 5 Personalised Online Learning Resource Delivery

In this chapter, we combine high-order feature interactions and the attention mechanisms to refine a recommender system, the proposed model combines several advantages from different state-of-the-art recommender systems and offers them in a smooth one-stop manner. This model enables automatic exploration of high-order feature interactions, differentiating the important degrees for different features, and mining latent features from the original input.

5.1 Background of the Use Case

The micro learning service aims to provide personalised small size learning materials in real-time. The learning material can be just some knowledge points [Lin, et al., 2019]. The online knowledge sharing service is one representative online learning service of the idea of micro learning. Hence, the investigation of the recommender system discussed in this chapter is mainly based on the application background of online knowledge sharing service. In the online knowledge platform, online users post and answer questions from various disciplines in a knowledge-sharing platform (like Quora[10] and Stackoverflow[11]), and the platform engages users with new questions or topics based on their profile and historical activities. However, the plethora of user interests and backgrounds could easily result in massive volumes of options and induce disengagement, i.e., producing questions that have the opposite effect. Hence, an online knowledge-sharing platform has to rely on a sophisticated recommender system to filter out irrelevant information to truly create a personalised learning service.

An effective recommender system needs to handle and merge different types and formats of information from the users' profiles and historical activities and resource profiles. Higher-order feature interaction (combination) is also crucial for good performance [Song, et al., 2019]. However, manually generating high-order feature interaction requires strong domain background. It is very time-consuming and labour-intensive, making it impractical for the large-scale online system, in the

[10] https://www.quora.com/
[11] https://stackoverflow.com/

context of big data. Furthermore, different features have various importance levels for a personalised recommendation task [Huang, et al., 2019]. How to precisely distinguish the importance differences of different features for a specific user is also vital for a personalised web-based learning service. Conventional recommendation strategies such as simple collaborative filtering and content-based filtering [Pazzani, 1999] are no longer adequate to handle massive, complex, dynamic data due to their drawbacks in scalability and modelling higher-order features.

5.2 Model Design

In this research, we aim to effectively combine these functionalities: mining and generating high-order feature interaction, distinguishing the importance differences of both implicit and explicit features, and maintaining the original input information in a single network. To this end, we propose a new deep cross attention network (DCAN) model for the recommendation task of the online knowledge sharing service. The input of the model contains both user-side and question-side information, and the embedding layer maps such information into a low dimensional space. The embedding vectors are then passed into the DNN network and crossing network separately for mining latent information and high-order feature interactions. The processed results are fused together, and an attention network is used to distinguish the importance differences of different features. Finally, the output layer is used to make predictions with weighted features. In this section, we firstly propose three hypotheses which might be significant to personalised micro learning service. The design of the model and technical details of each component is then presented and discussed.

5.2.1 Hypotheses

In this study, the proposed model is designed based on the following hypotheses:

1. High-order feature interaction is vital to further improve the performance of a recommender system which used for a web-based big data application. Low-order feature interaction cannot sufficiently mine and model the underlying complex feature interactions for informal learning service.
2. The features involved in a learning platform have different important degrees. Precisely differentiate the feature importance is vital to a personalized learning service, and it can further improve the recommendation results.

3. The proposed model also holds moderate efficiency. For a web-based learning service in the big data era, efficiency is also an important indicator.

For example, the proposed model manages to recommend the question to the user that he/she might be interested in. The recommended question given by the system is about machine learning; the difficulty level of this question is entry-level. And we have two users with the following features:

User_1: (Interested topic: computer science, Occupation: student, Gender: male, Age:23, Location: Australia)

User_2: (Interested topic: computer science, Occupation: research fellow, Gender: male, Age 32, Location: China)

For the example, the proposed model should effectively and automatically generate meaningful feature combinations such as *(Interested topic, Occupation)* and *(Gender, Age)*, distinguish the importance difference between different features such as for a given question the feature *(Interested topic, Occupation)* might be more important than the feature *(Gender, Age)* and decide that the *User_1* might be more interested in this question.

5.2.2 Model Architecture

Based on the above hypotheses and one previous proposed initial idea [Lin, et al., 2020], the general architecture of our model is shown in Figure 5.1. Our proposed model contains three significant networks: a cross-network for exploring feature interactions, a deep network for mining latent information, and an attention network for distinguishing the importance of different features. The input of the model is high dimensional vectors that contain both user and item relevant information. The output of the model are decimals range from 0~1 indicating how much a user will be interested in the given question.

5.2.2.1 The Embedding Layer

For a recommendation task, the input contains many highly sparse categorical data, such as the genre describing the discipline of the educational resource, which may be multi-valued (for example, the subject of '*machine learning*' could belong to both disciplines of '*mathematics*' and '*computer science*'). In our proposed model, we apply an embedding layer to reduce the dimensionality and sparsity of the raw data. The raw input contains the history of interaction between the user and the online resource and the side information of the user and the online resource. The embedding

operation can not only reduce the computational workload, but also boost the model performance. This process can be formulated as Equation (5.1):

$$X_{embed,i} = W_{embed,i} X_i \tag{5.1}$$

Where $X_{embed,i}$ is the embedding result of the *i*-th categorical feature, $W_{embed,i}$ is the embedding matrix that maps the *i*-th original categorical feature into the low dimensional space, and X_i is the i-th feature.

5.2.2.2 The Cross Network

The cross-network used in this study is based on the method proposed in [Wang, et al., 2017]. The cross-network is used to automatically generate the high-order feature interactions. Such network consists of several layers, and each layer is an operation of feature interaction. For each interaction layer, the operation of feature crossing can be simply formulated as Equation (5.2):

$$X_{l+1} = X_0 X_l^T W_l + b_l + X_l \tag{5.2}$$

Where X_l is the output of the *l*-th crossing layer, W_i and b_i are weights and bias parameters of each crossing-layer. As demonstrated in [Wang, et al., 2017], such special structure of network can increase the interaction degree as the network goes deep, with the highest *n+1* polynomial degree of the *n*-th layer. Moreover, regarding the efficiency of the network [Wang, et al., 2017], the time and space complexity are both linear in input dimension.

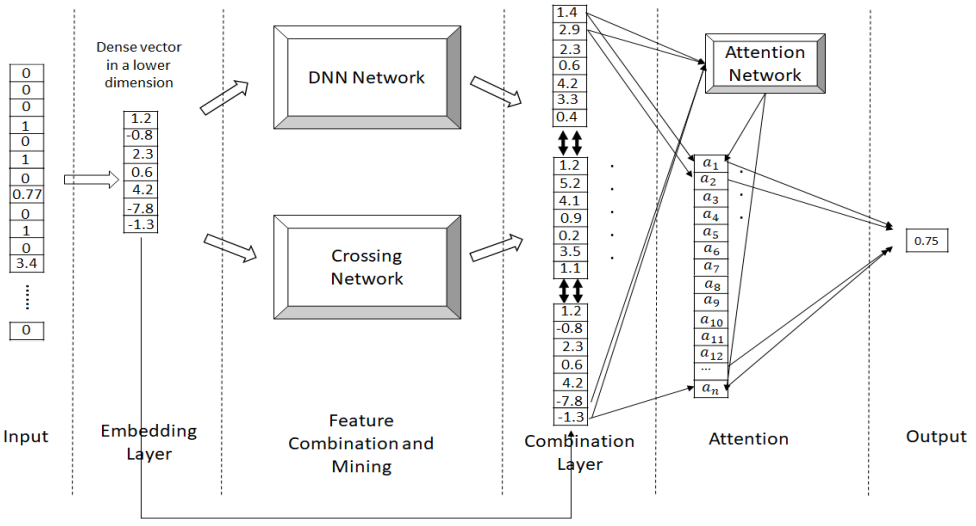

Figure 5.1 The Overall Network Structure of the Proposed Cross Attention Boosted Recommender System

5.2.2.3 The Deep Network

A conventional fully connected neural network (multi-layer perceptron) is used in the proposed model as the deep component for simplicity and generalisation. The deep network implicitly captures the latent information and feature combinations. Each layer of the DNN network can be formulated as Equation (5.3):

$$h_{l+1} = f(W_l h_l + b_l) \tag{5.3}$$

Where h_l denotes the output of the *l*-th layer of the deep component, $f(\cdot)$ is the activate function, where ReLU is used in this study. W_l and b_l are parameters of the *l*-th layer of the deep network.

5.2.2.4 The Residual Connection and Attention Network

Before providing transformed information to the attention network, the original input information is continuously added to the output of the deep network and the cross-network by using residual. This aims to maintain the original input information, which might suffer information loss after going through several layers of the neural network. The residual connection used in this research connects the output of one earlier layer to the input of another future layer several layers later.

An attention network is applied right after the combination layer to interpret the important difference of various features. The attention mechanism can be formulated as Equation (5.4) and (5.5):

$$a_i' = ReLU(WX_i + b) \tag{5.4}$$

$$a_i = \frac{\exp(a_i')}{\sum_i \exp(a_i')} \tag{5.5}$$

where, both W and b are the model parameters. The attention score is calculated through Softmax function. The calculated attention scores are projected back to the output of the combination layer. The process can be formulated as Equation (5.6):

$$X_s = a_i WX \tag{5.6}$$

Where, a_i is the calculated attention score, W is the weight adopted in the network, and the X is the output of the former combination layer, and X_s is the final value after attention mechanism is applied. The demonstration of the attention operation is shown in Figure 5.2. Where $\boldsymbol{f_1}$ to $\boldsymbol{f_n}$ are the latent features passed into the attention network, the outputs of the network are attention scores for latent features with Softmax function. Lastly, each score is assigned to the feature by Hadamard product.

5.3 Experiment and Analysis

In this section, we compare our proposed model with several state-of-the-art recommendation strategies.

5.3.1 Evaluation Metrics

In the experiment, we used the Area Under Curve (AUC) as the main criteria to evaluate the performance of each model. The proposed is a binary classifier which predicts whether a user will be interested in a given question, and the AUC can measure the capability of a model in distinguishing two labels. The calculation of the AUC used in this study is calculated as Equation (5.7):

$$\text{AUC} = \frac{\sum_{i \in positiveClass} rank_i - \frac{M(1+M)}{2}}{M \times N} \tag{5.7}$$

Where M and N are the number of positive and negative samples respectively, $rank_i$ is the location of the i-th sample. We also used mean square error (MSE) and binary cross entropy to reflect the errors that made by each model. The binary cross entropy used in this study is formulate as Equation (5.8):

$$\text{H} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\big(p(y_i)\big) + (1 - y_i) \log(1 - p(y_i)) \tag{5.8}$$

5.3.2 Dataset

The dataset used in the experiment contains 10 million (question, user) pair, which is collected from Zhihu[12]. The user here stands for an online learner who is or is not a participant in a certain question. This dataset also contains other types of side-information about the users and the questions, such as the answers to the question, categorical information (such as gender) about the user, and the user's learning interests. Moreover, the distribution of the samples is unbalanced (the ratio of negative

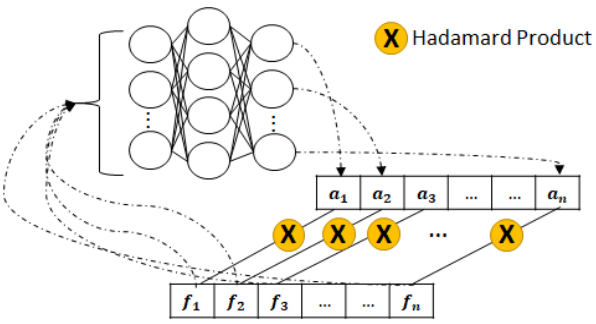

Figure 5.2 Visualization of the Attention Operation

[12] https://www.zhihu.com/

samples to positive samples is around 4), which reflects the real and typical online application scenario, where for most users, only a small amount of questions they would like to answer due to various reasons, such as 'pedagogical lurking'. A negative sample stands for recommending a question to a user, but the user does not participate in any learning activities; while a positive sample stands for recommending a question to a user and the user participate in a certain learning activity which could be answering the question or commenting the answers given by other users. As discussed in many pedagogical studies [Dobozy, 2017, Beaudoin, 2002, Dennen, 2008], for the online learning service, it is very difficult to enable learners to interact with each other like offline learning, even if the online learners have great and similar interests in the current learning session.

5.3.3 Baselines

AutoInt [Song, et al., 2019], FM, DeepFM, Deep&Cross Network (DCN), Attention Factorization Machine (AFM) are used as baselines in the experiment. The overall architecture comparison of AutoInt, DeepFM, AFM, and DCN is shown in Figure 5.3. Each of these models contains an embedding layer, a feature interaction layer and uses the Softmax function to make a prediction. The main difference among these models is the choice of techniques for feature interaction, where multi-head self-attention is used in AutoInt, the combination of the FM and the DNN is used in DeepFM, the combination of FM and simple forward attention network is used in AFM, and combination of the cross-network and the FM is used in DCN.

5.3.4 Experimental Setup

All the models involved in the experiments are implemented using PyTorch [Paszke, et al., 2019]. Each categorical feature is represented as an embedding vector with six dimensions. All the non-linear transformations are activated by ReLU except the output layer. The output is activated by the softmax function. All baselines are implemented strictly follow the suggestions and guideline of their original research. The early-stop mechanism is applied to all models involved in this experiment in order to prevent overfitting. Ten-fold-cross-validation is applied in the experiment.

5.3.5 Experiment Results and Analysis

The comparative experiment result is shown in Table 5.1 and Figure 5.4, which illustrates the overall

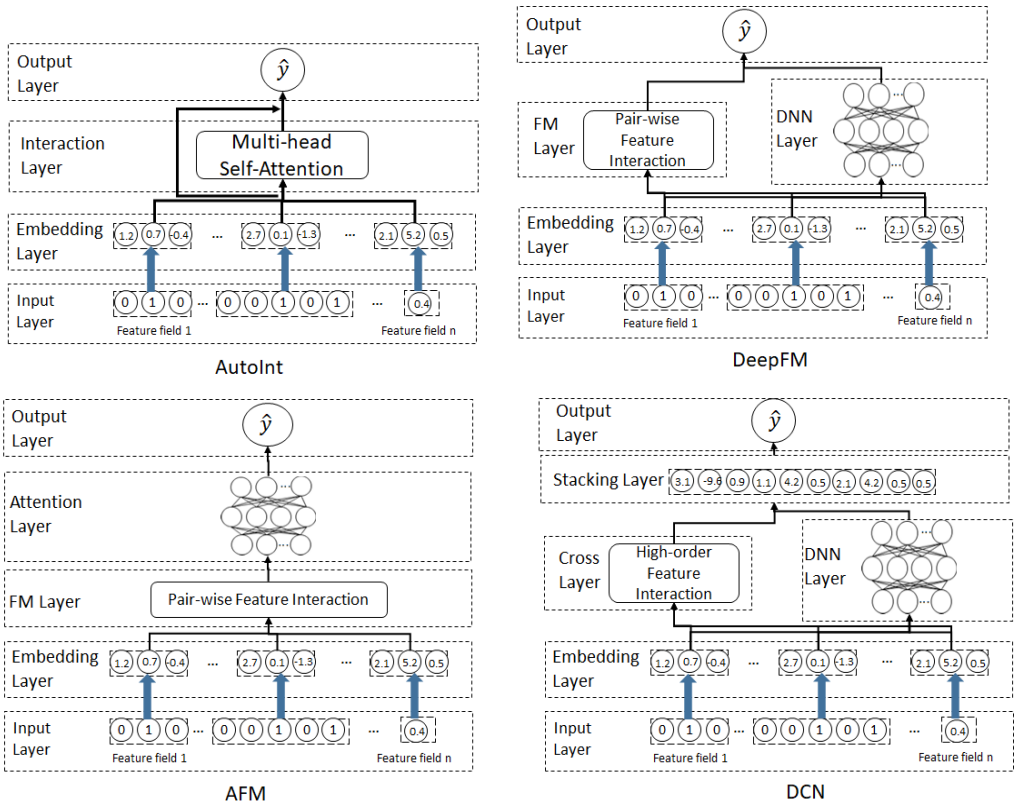

Figure 5.3 Overall Model structure of AutoInt, DeepFM, AFM, and DCN

performance of each model based on three different criteria, AUC, MSE, and binary cross-entropy. According to the results, we can easily get the following three conclusions.

5.3.5.1 The Importance of the High-order Feature Interaction

According to the results in Table 5.1, we can clearly see that the AUC scores of FM and AFM are the lowest ones, and the MSE and binary cross-entropy values of these two models are the highest. These two models are the only two of which involves up to second-order feature interactions. Other baseline methods and our proposed model all involved high-order feature interactions, even though the ways of feature interaction are different. Hence, we argue that the high-order feature interaction (complex feature combination) does reflect how online learners make their decisions and involving the high-order feature interaction is useful and necessary to the large-scale web-based learning recommendation task. This finding proves the first hypothesis that made in the previous section.

5.3.5.2 The Significance of the Attention Mechanism

Another conclusion we can get from Table 5.1 is the models (AFM and our proposed model) that involve the attention mechanism have higher AUC scores than the models which do not. One

Table 5.1 Experiment Results of Different Models

Model	AUC	MSE	Binary cross entropy
FM	0.6934	0.1243	0.4060
DCN	0.7603	0.1134	0.3690
AFM	0.6881	0.1255	0.4094
AutoInt	0.7613	0.1130	0.3679
DeepFM	0.7404	0.1128	0.3671
Proposed model	**0.7848**	**0.1071**	**0.3442**

possible explanation is AFM and our proposed model refine the results of high-order feature interaction via the attention mechanism. Such a result approves the second hypothesis that we made in the previous section. The main difference between the AutoInt and the proposed model is that they use different techniques to explore the feature interactions. According to the experiment result, we can see that our proposed model outperforms the AutoInt when handling the recommendation problem in the online knowledge sharing scenario.

5.3.5.3 The Efficiency of the Proposed Model

To investigate the third hypothesis that we made in the previous section, we also evaluate the computation efficiency of our proposed model and various state-of-the-art recommendation models. The result is also shown in Figure 5.4. The proposed model is in third place, outperforming the FM, AutoInt and DeepFM, and very close to the second one (AFM). However, the AFM does not involve high-order feature interactions. The most efficient model is DCN, which only takes around 195 seconds under our experiment setting, while our proposed model takes 261 seconds on average for each training epoch. As the network architecture of our proposed model is extended from DCN and much more complex than other baselines, considering the improvements in recommendation performance and other models that have been verified their efficiency on real-world applications, the slight increase of the training time is reasonable and acceptable. The possible reason for this is many operations involved in the proposed model run simultaneously (such as the crossing network and deep network). Hence, we trust that with proper configurations, our model can reach efficiency requirements.

5.4 Summary

In this Chapter, we refine an existing recommender system [Lin, et al., 2020] using the attention

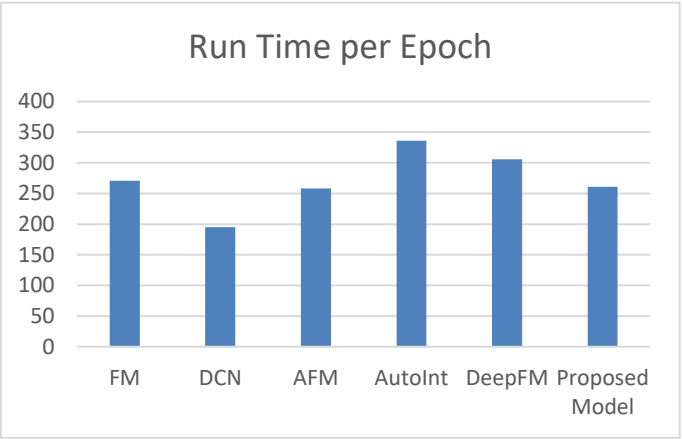

Figure 5.4 Efficiency Comparison of Different Models in Terms of Rum Time (s/epoch)

mechanism together with high-order feature interaction methods to boost the performance of a web-based knowledge-sharing service. By comparing with the state-of-the-art recommender system, we confirm three hypotheses about the proposed model: 1. The involved high-order interaction is meaningful and can help further boosting recommendation performance. 2. Features used in the recommender system have different degrees of importance. The attention mechanism can better distinguish such difference comparing to the conventional weighting method. 3. Even though the structure of our model is more complex than the baselines, it still shows acceptable running efficiency. The experiment results clearly demonstrate that our model has the potential in handling complex online learning recommendation problem. More specifically, according to the experiment results with authentic online knowledge sharing data, the strengths of DCAN can be concluded into two points: 1.the proposed model can automatically mine and generate high-order feature interactions in both explicit and implicit ways; 2. the proposed model can further distinguish the importance differences of different features.

For future directions, it is worth further exploring the recommendation strategy for online learning service. It is worth investigating how to precisely represent, model and integrate chronological or temporal factor in the recommendation task. As highlighted in [Zhou, et al., 2019], with the changing external environment and the internal cognition, a user's interest might evolve over time. Especially for the online informal and non-formal educational activities, many factors are dynamic, such as learning interest drifting and the changes in knowledge level.

Chapter 6 Student Dropout Rate Prediction

In this chapter, we propose a double-tower-based framework for dropout prediction. It separately models the macro and micro information from the learner's historical interactions with the course contents. This chapter also demonstrates the design of a Convolutional Neural Network (CNN)-based model for effectively mining time-series information from the online learner's sequence of activity records.

6.1 Research Questions

As a course is made up of several instructional videos, the interactions between a user and a course can be regarded as a sequence of interaction logs with such videos. In general, both the course structure and the learner's interaction record are in the form of sequential patterns. Whether a learner will drop out from an enrolled course can be predicted based on the sequence of historical online learning behaviours he/she has taken. This suggests the need for using a temporal modelling method to mine the time-series patterns, i.e., discovering patterns from the learner's historical activity record. Hence, to better predict the dropout rate based on the learner's historical activity records of the course, it is also vital to mine the hidden time-series patterns. A double tower-based deep learning framework with a carefully designed CNN-based micro component is proposed in this chapter to tackle the above research motivation. The contribution of our work mainly tries to answer the following research questions and address the related problems:

1. Does the user's interaction log record contain any useful time-series patterns for predicting learner dropout rate?
2. For learner behaviour modelling like dropout rate prediction, is it necessary to use different modelling components to handle different granularities of information?
3. If the data contains the latent time-series patterns, what is an effective method to model such patterns?

6.2 Model Design

This section describes the architecture of our proposed double-tower framework from a high-level perspective at first. Then, the design details of the framework are presented and discussed. Subsequently, the relevant loss functions used in the training steps are explained.

6.2.1 The Architecture of the Double-Tower Framework

Inspired by the prior exceptional work in the recommender systems [Cheng, et al., 2016], the proposed framework is in a double-tower structure containing two intelligent components, one for modelling the macro information and another for modelling the micro information (see Figure. 6.1). The design of the framework structure is based on the idea of 'separate and conquer', which aims to separate different types of information and use the most suitable model to conquer each type of information. The raw input comes from the user's interaction logs and the courses' profile, which contain a mixture of both micro and macro information (indicated in blue and red colour in Figure 6.1, respectively). Each component takes and handles different information and produces an intermediate result solely bases on one type of information (micro or macro). At last, two intermediate results are summarized through a regression function and used to produce the final prediction.

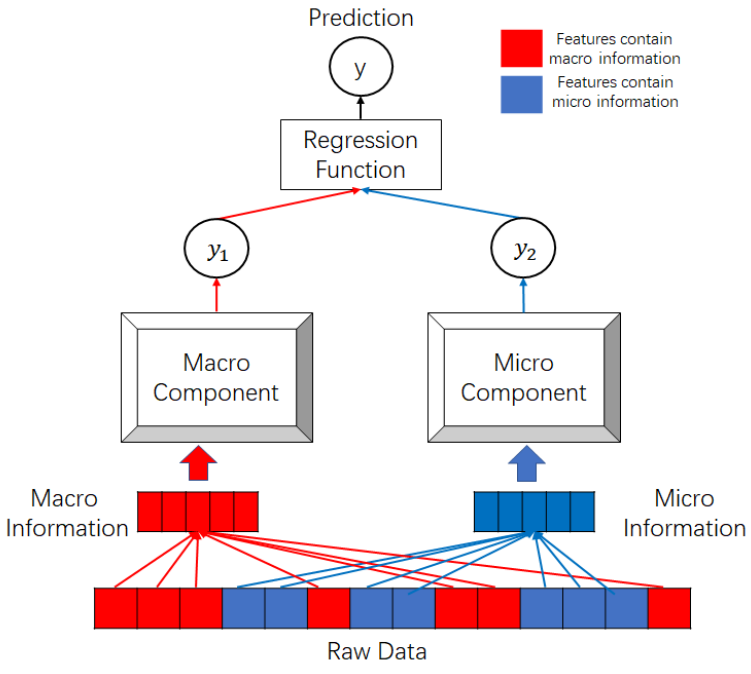

Figure 6.1 The Overall Network Structure of the Proposed Double-Tower Framework

6.2.2 Problem Formulation

To clearly formulate the problem of the learner dropout prediction, we firstly present the following definitions.

6.2.2.1 Objective

Given the input with different types and granularities of information, the goal of the proposed framework is to predict whether a learner will drop a given online course. Mathematically, such a task can be regarded as a binary classification problem. Defining the ground truth as $y \in \{0, 1\}$ (where 1 stands for dropout and 0 for not dropout), and the prior knowledge of course profile and user's historical learning records as *x*. Given the input signal *x* and the target output *y*, the prediction process is notated as $\mathcal{F}$and can be simply formulated as Equation (6.1):

$$\mathcal{F}(x) \Longrightarrow \mathrm{y} \tag{6.1}$$

6.2.2.2 Micro Information

In this study, the term of micro information refers to the interaction details (i.e. watching times, video watching start time, and video watching end time) of a certain user u_i with a course $\boldsymbol{c}$; it is formulated as $\boldsymbol{L}_i$ (u_i, $\boldsymbol{c}$). As one course may contain several instruction videos, the above definition can be further notated as $\boldsymbol{L}_i$ (u_i, $\boldsymbol{c}$)$\in \mathbb{R}^{nxm}, \boldsymbol{c} = \{v_1, v_2, \dots, v_n\}$; where *n* is the number of the videos (belonging to the course $\boldsymbol{c}$) that the user u_i has already interacted with, *m* is the dimensionality of the feature space. The features notated as $f_1 \sim f_m$ represent the interaction detail between the user u_i and a video v_j. Specifically, given a course $\boldsymbol{c}$, $\boldsymbol{L}_{i,j}$ is the interaction record between a certain user u_i and a certain video v_j, and $\boldsymbol{L}_{i,j}(u_i, v_j) \in \boldsymbol{L}_i(u_i, \boldsymbol{c})$. Inspired by the research of pattern recognition from computer vision, in this study, the interaction between a user and a course is organised and represented in the form of a 'figure'. The organisation and relationships between the user-video interaction and the user-course interaction are thus shown in Figure 6.2. As such organisation of the information may contain latent time-series pattern, a deep learning-based model is used to mine the underlying features and make predictions.

6.2.2.3 Macro Information

The macro information refers to the high-level, general information of a certain course. This includes basic profile information of this course (i.e. description, discipline, and the number of videos) and

descriptive statistical information about all learners' interaction history with the course (i.e. the course popularity and the number of enrolled students). For the interaction between a certain user u_i and a course $\boldsymbol{c}$, the macro information is formulated as $\boldsymbol{G}_i(u_i, \boldsymbol{c}) \in \mathbb{R}^h$, where h is the dimensionality that used to represent the macro information. As the macro information only includes high-level general information without complex patterns, to reduce the time and space complexity, a less complex model (e.g., logistic regression) will be used to process them.

Together with the definition of the objective and micro information, the dropout prediction task can be further defined as: given the macro information $\boldsymbol{G}_i(u_i, \boldsymbol{c})$ and the micro information $\boldsymbol{L}_i(u_i, \boldsymbol{c})$ between the user u_i and the course $\boldsymbol{c}$ as the input signal, our goal is to predict whether the user u_i will drop out from course $\boldsymbol{c}$ in the future. The proposed framework is to learn the following Equation (6.2) $\mathcal{F}$:

$$\mathcal{F}\big(\boldsymbol{G}_i(u_i, \boldsymbol{c}), \boldsymbol{L}_i(u_i, \boldsymbol{c})\big) \Longrightarrow y_{(u_i, \boldsymbol{c})} \tag{6.2}$$

6.2.2.4 Convolutional Network with Fixed Kernel Width

Inspired by the network structure proposed in the prior work reported in [Kim, 2014], we have also carefully designed a CNN-based network to capture the micro information, which may contain valuable time-series patterns. The interaction 'figure' is scanned by multiple kernels with a fixed-width *m*, which equals to the feature size (the dimension size). Hence, each convolutional operation summarizes k_i pieces of successive learning activities where k_i equals to the kernel height. As illustrated in Figure 6.3, two kernels, denoted in red and blue, with heights k_1 and k_2, summarize successive learning activities with different time length k_1 and k_2. Next, the extracted time-series patterns are further processed and summarized by following convolutional and pooling operations. A

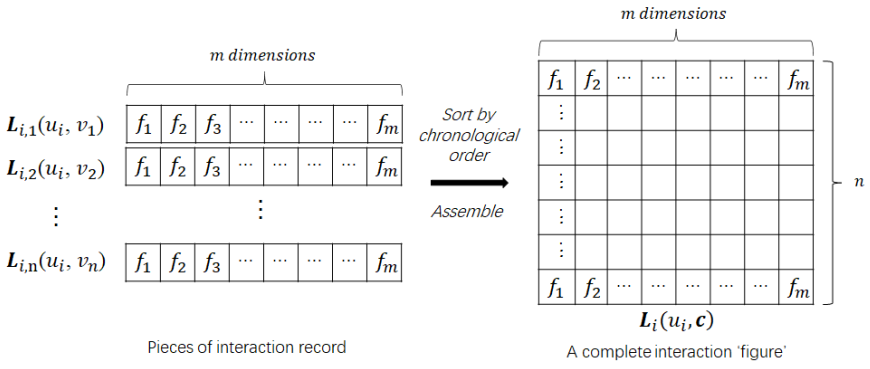

Figure 6.2 The Organisation Detail of an Interaction 'Figure'

fully connected layer is used to produce the intermediate result for the micro component. For the comprehensive evaluation of the time-series modelling capability, in our experiments, the mainstream solutions such as LSTM, GRU and their variants were also implemented and compared with the proposed CNN-based model. The details will be discussed in Section 6.2.3.

6.2.3 Separate and Joint Training Strategies

As shown in Figure 6.1, the intermediate results produced by the micro component and the macro component are combined using a weighted sum before being passed into the regression layer for the final prediction. This processing step is formulated as Equation (6.3).

$$P(Y = 1|x) = \sigma(W_{micro} r_{micro} + W_{macro} r_{macro} + b) \tag{6.3}$$

where r_{micro} and r_{macro} are the intermediate results from micro and macro components, respectively; W_{micro} and W_{macro} are two different weights; b is the bias value; and σ represents regression function.

Note that in this framework, the micro and macro components can be trained in two different ways: joint mode or separate mode. As its name implies, in the joint mode, two components are trained jointly. It is an end-to-end training process with all of the parameters optimized simultaneously. Conversely, in the separate mode, two components are trained separately without knowing each other. When using the seperate mode, after training, the two intermediate results can be directly combined by applying weighted means or be combined through an additional regression layer, same as Equation (3). As for training the joint mode, an additional loss function can be used to add

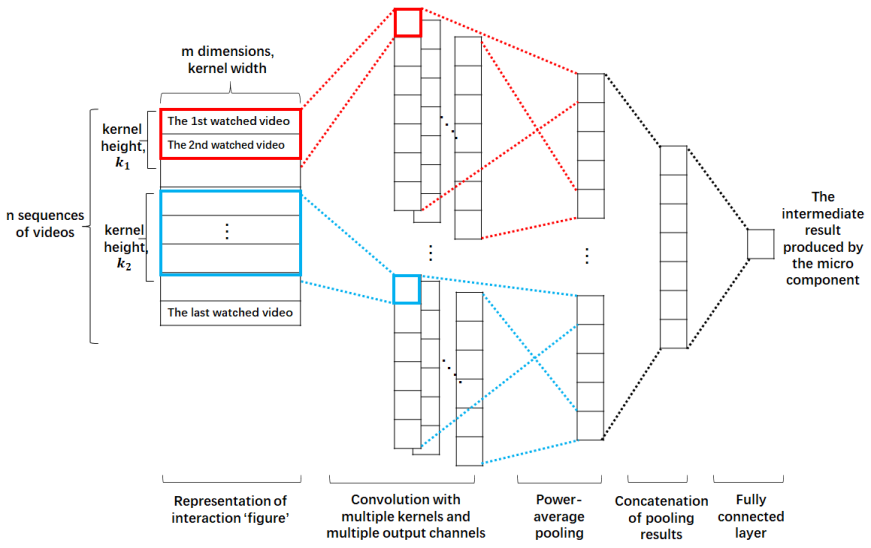

Figure 6.3The Network Structure of the CNN-based Micro Component

constraints for the intermediate output of each component. As formulated as Equation (6.4) and (6.5), it is to ensure both components focus on predicting the learner's dropout rate.

$$Loss_l = -\sum_{i=1}^{n} l_{r_i} logy_i + (1 - l_{r_i}) \log(1 - l_{r_i}) \quad (6.4)$$

$$Loss_g = -\sum_{i=1}^{n} g_{r_i} logy_i + (1 - g_{r_i}) \log(1 - g_{r_i}) \quad (6.5)$$

where l_r_i and g_r_i are the intermediate outputs of the micro and macro component, respectively. The y_i is the ground truth label for the i-th sample. The final loss for the joint mode was calculated in Equation (6.6).

$$Loss = Loss_l + Loss_g + Loss_f \quad (6.6)$$

where $Loss_l$ and $Loss_g$ are the loss values for micro component and macro component, respectively. $Loss_f$ is the final loss of the entire double-tower framework.

6.3 Experiment and Analysis

The details of the conducted experiments are discussed in this section, including details of the used dataset, evaluation metrics, involved baselines, relevant experiment settings and the comparative analysis of the results.

6.3.1 Dataset

The dataset used in this experiment was collected from XuetangX[13], the largest MOOC platform in China launched by Tsinghua University in 2013. Up to now, this platform has hosted more than three thousand high-quality courses and has more than 58 million registered online users. The data used in the experiment was extracted from the public data repository MOOCCube[14], which hosted 706 online courses, more than 30 thousand videos, and about 200 thousand users. We used data from all available courses and videos, but we left out the information about the concept, taxonomy, and paper information of the dataset, as this was beyond the scope of this study. For details of the statistic information about MOOCCube and the explanation of the dataset, please refer to the original work reported in [Yu, et al., 2020].

6.3.2 Evaluation Metrics

[13] https://www.xuetangx.com/
[14] http://moocdata.cn/data/MOOCCube

An intuitive approach to directly measure the overall performance of a binary classification model is by using accuracy (ACC). This is the proportion of the correct predictions out of the total number of predictions, as formulated in Equation (6.7):

$$ACC = \frac{total\ number\ of\ correct\ prediction}{total\ number\ of\ prediction} \quad (6.7)$$

It should be noted that the accuracy paradox might occur [Valverde-Albacete, et al., 2014], especially for the imbalanced distributed dataset. Hence, for a comprehensive evaluation, we also involved some other evaluation metrics discussed as follows.

Another evaluation metric used in the experiment is the F-measure, which is the harmonic mean of Recall and Precision. It is worth mentioning that due to the trade-off between Recall and Precision, high Recall or Precision cannot directly and correctly indicate a good model. Hence, F-measure will be a better option for evaluating the binary classification results. Specifically, F1-score was used in the experiment.

To achieve a more comprehensive evaluation result, the Area Under Curve (AUC) score is also used as an evaluation metric in the experiments. Known as the area under the Receiver Operating Characteristic (ROC) curve, AUC reflects the distinguishing ability of positive and negative labels for a model.

6.3.3 Baselines

In the experiments, to comprehensively investigate the learner dropout prediction problem for the MOOC learning environment and demonstrate the effectiveness of our proposed solution, different models were used to construct the proposed double-tower framework. For the macro component, we applied:

1. LR: logistic regression model. It is one of the most straightforward and efficient models that requires very little computation power. LR can also give direct inference about the importance of the involved features.
2. GBDT: gradient boosting decision tree. GBDT and its variants have been proved to be a good choice to handle machine learning tasks in the industry area and various data science competitions. GBDT has several advantages, such as flexibility, less data pre-processing required, and the ability in handling missing data.

For the micro component, we implemented and compared the following models:

1. The proposed CNN-based neural network.
2. (Bi-)GRU: (bi-directional) gated recurrent unit neural network. (Bi-)GRU is more efficient than (Bi-)LSTM, with less training time.
3. (Bi-)LSTM: (bi-directional) long-short-term-memory neural network. Comparing to (Bi-)GRU, (Bi-)LSTM shows better performance when mining the long sequence pattern.
4. MLP: multi-layer perceptron neural network. MLP is the most intuitive and simple network, it is frequently used as the benchmark when the experiment involves comparing different deep learning or neural network models.

6.3.4 Experimental Setup

All neural networks are implemented using the PyTorch framework [Paszke, et al., 2019]. To ensure that the objective functions work properly, all continuous features used in selected neural networks are scaled using the z-score normalization strategy. The dimension number of the hidden layer for GRU and LSTM is set to 64, and 32 for Bi-GRU and Bi-LSTM. For the CNN model, three different kernels (height 3, 4, and 5) are used with 32 output channels. Power-average pooling with Power 2 is used in the CNN network. The input of the MLP is the concatenation of shuffled interaction records. We use the shuffled data in MLP to verify the existence of the time-series pattern in the micro information. The categorical features are converted into dense embeddings. ReLU is used as the activation function for all non-linear transformation layers except for the output layer, which is activated by a Sigmoid function. All the other settings strictly followed the guidance of the original work of each model or the default settings in the used frameworks and libraries.

6.3.5 Experimental Results and Discussion

6.3.5.1 The Existence of the Time-Series Pattern

To find out whether the interaction log contains any latent time-series information, in the experiment, we firstly apply several models with/without time-series modelling mechanisms to handle the micro and macro information. Specifically, LR, GBDT and MLP are the models that do not involve any specialized time-series modelling mechanisms. The GRU, LSTM and the proposed CNN-based network are the ones that contain time-series modelling mechanisms. According to the comparison results (see Table 6.1), it is comfortable to find that the models with the time-series modelling

mechanism outperform the ones without it. This finding suggests the existence of the latent time-series pattern in the online learning interaction log, which answers the first research question in Section 6.1. As the macro information only contains the category and demographic information (such as subject popularity, number of courses involved, and subject discipline), we can further conclude that the time-series pattern exists in the micro information. Hence, to achieve better performance, we should use the model with time-series pattern modelling ability when handling the micro information.

6.3.5.2 The Effectiveness of the Double-Tower Framework

By comparing the results from Table 6.1 and Table 6.2, we can conclude that the double-tower framework in either joint mode or separate mode outperforms the single model. Using different components to handle different types of information, respectively, can further improve the model performance. Solely using a single model to process all types of information is difficult to find the optimal solution. This observation answers the second research question aforementioned in Section 6.1. The diversity of the information in an online learning platform like MOOC requires us to involve different models for better system performance. Moreover, when closely looking at the results in Table 6.2, we can also find out that the effectiveness of the double-tower framework is determined by the model used in each component and the training mode. The analysis of this result will be discussed in the remainder of this section.

6.3.5.3 The Specificity of the Time-Series Information in MOOC

When comparing the time-series modelling ability of the proposed model with (Bi-) LSTM and (Bi-) GRU, the results clearly suggest that the proposed CNN-based model surpasses the rest, either in

Table 6.1 Comparison of Single Model

Models	ACC	F1	AUC
LR	0.7727	0.7037	0.8382
GBDT	0.7967	0.7293	0.8619
Proposed	**0.8292**	**0.7859**	**0.8948**
GRU	0.8211	0.7730	0.8918
LSTM	0.8142	0.7560	0.8892
MLP	0.8032	0.7462	0.8739

single model comparison (see Table 6.1, highlighted in bold) or to be compared as a component in the double-tower framework (see Table 6.2, highlighted in bold). Specifically, the framework with a GRU-based micro component outperformed the one with the LSTM-based micro component. Bi-directional GRU or LSTM did not show any remarkable improvements, and sometimes they were even worse (in the separate mode). Similar results can be observed in Table 6.2. Hence, we can conclude the result of model performance as CNN > GRU > LSTM, and GRU/LSTM ≈ Bi-GRU/Bi-LSTM[15].

The above finding answers the last research question mentioned in Section 6.1, that comparing to the mainstream time-series modelling solution such as LSTM and GRU, our model is more suitable for an online learning scenario. This can be ascribed to the specificity of the time-series information in the online educational activity. For online learning, the relationships between information sequences are relatively weak.

Based on the mathematical concepts of LSTM [Gers, et al., 1999] and GRU [Cho, et al., 2014], LSTM has a more complex structure, and tends to model longer sequences than GRU. In other

Table 6.2 Comparison of the Separate Mode and Joint Mode

Models	Separate Mode			Joint Mode		
	ACC	F1	AUC	ACC	F1	AUC
LR + Proposed	**0.8363**	**0.7875**	**0.9004**	**0.8566**	**0.8223**	**0.9194**
LR + GRU	0.8259	0.7755	0.8940	0.8449	0.8109	0.9143
LR + Bi-GRU	0.8220	0.7694	0.8915	0.8472	0.8120	0.9178
LR + LSTM	0.8187	0.7601	0.8895	0.8377	0.7908	0.8972
LR + Bi-LSTM	0.8172	0.7639	0.8858	0.8396	0.7928	0.9013
GBDT + Proposed	**0.8423**	**0.8016**	**0.9093**			
GBDT + GRU	0.8395	0.7988	0.9095			
GBDT + Bi-GRU	0.8386	0.7962	0.9084			

[15] Marks '>' and '≈' are used as shorthand to represent outperformance and similar performance, respectively.

words, the LSTM model tries to encode more long-term and short-term information than GRU, but for the recording of online learning activity, the dataset does not actually contain much relevant time-series information. Similarly, whilst bi-directional LSTM or GRU tries to model more complex sequential information, the result suggests that the dataset does not contain much complex time-series information.

In most cases where LSTM and GRU outperform the CNN model, the datasets always contain 'rich' time-series information. Such sequence of information is strictly composed of certain rules or domain-specific requirements or features, such as the grammar rules in NLP, and genetic patterns in the DNA sequence.

However, for the activity records of online learning, the connection between adjacent activities is weaker than the relationships between words in the NLP task. Each video is a relatively independent unit, which contains a certain amount of complete knowledge points. The interactions with videos can be relatively random. Although the learning records in MOOC do contain some useful time-series information, which could be useful for dropout prediction, such information does not contain the obvious long-term pattern. Therefore, for mining or modelling sequence of learning activities, it is suggested to choose a model with better generalization ability and less constraint.

6.3.5.4 Comparison of Effectiveness and Efficiency between Two Training Modes

When comparing the left part of the first five results with the right part in Table 6.2, we can see that training micro and macro components jointly produces better results than training them separately. We ascribe this improvement to sufficient information exploring in the joint mode. As an end-to-end framework, all the information is exposed to the framework and exchanged between two components at the same time. In the experiments, we also monitored the training time in order to compare the efficiency of the proposed framework in different settings. Table 6.3 records the running time (seconds) per training epoch and the number of epochs a training process needs to reach the optimal point[16]. It is worth mentioning that when using the separate mode, two components can be trained simultaneously. In addition, the macro component requires much less training time than the micro component as it has a simpler structure. Hence, we had merely measured the running time of the micro component for the separate mode.

[16] The optimal point does no refer to the point that model training terminates. As the neural networks were initialized by random seed every time, it is an approximated epoch number that the model required to get close to the final state.

The frameworks with the proposed CNN-based component requires the least time for both joint mode and separate mode (see Table 6.3, the first and fourth row). The frameworks with the LSTM component took longer time in each training epoch comparing to the ones with GRU. Notably, such a result corresponds with the mathematical definitions [Cho, et al., 2014] of these two models, where GRU is much simpler and lightweight than LSTM. Furthermore, the joint mode required a little more time for each training epoch but less training epoch to reach the optimal point (see Table 6.3, the comparison of the first three results and the last three results). This also corresponds to the mathematical derivation from which the joint mode has involved more parameters in the training process and has higher time complexity. As the joint mode required fewer training epochs, it is also viable to conclude that the additional loss functions (Equation 6.6) used in the joint mode have forced the two components to efficiently exchange information flow during the training process. Such internal information exchange also exposed additional information to both components.

6.3.5.5 The Implication of Two Training Modes

As for the implications, the joint mode is an end-to-end training process, which is a more efficient training strategy. Compared to the separate mode, it requires less training time and fewer epochs to find the optimal solution, as discussed before.

However, in some cases, these two components cannot be trained simultaneously. For example, in some high-performance conventional models, such as the random forest, they are very difficult to be trained together with a neural network because they use quite different optimization strategies. To better demonstrate such a situation, GBDT was also used as another type of macro component, which was difficult to implement under the PyTorch framework. According to Table 6.2, the

Table 6.3 Comparison of the Framework Efficiency

Models	Epoch number	Time (s/epoch)
LR + Proposed (J)	15	287.12
LR + GRU (J)	15	320.35
LR + LSTM (J)	15	345.04
LR + Proposed (S)	25	286.36
LR + GRU (S)	25	294.97
LR + LSTM (S)	25	327.27

framework using GBDT as the macro component outperformed those having used LR as the macro component in the same setting. Empirically, one of the two components may reuse an existing trained model. This will allow us to only train one new model for another component. Undoubtedly, in these situations, the separate mode is more efficient than the joint mode. In summary, the joint is more efficient, but the separate mode is more flexible than the joint mode.

6.4 Summary

In this study, we investigate and analyse the learner dropout prediction problem for the MOOC platform. In order to deal with different types of information, we propose a double-tower framework. We design a CNN-based network to model the time-series information from users' successive learning records. In the experiments, the proposed framework is comprehensively evaluated by applying different settings and combinations of components. The difference in time-series modelling approaches between dropout rate prediction tasks and the tasks from other disciplines (e.g. NLP and bioinformatics) are also discussed. The results have shown that the proposed framework with a CNN-based micro component outperforms all baseline models. The experiments demonstrate that, both the proposed double-tower framework and the CNN-based neural network are entitled with high effectiveness. This novel work also provides empirical implications for the practical usage of different training modes. Moreover, the proposed model shows effectiveness in both the joint mode and the separate mode.

In our future research, to ensure the model to be more robust, we will seek to integrate other state-of-the-art techniques, such as attention-based and residual networks.

Chapter 7 Generative Adversarial Network Based Optimization Strategy for the Micro Learning Recommender System

This chapter highlights and discusses advantages of GAN-based recommender system and gaps with micro learning service through comprehensive discussions and a pilot experiment. This chapter also proposes a novel adversarial training model to balance the trade-off between the false-negative rate and false-positive rate of the recommendation results and mine the implicit non-positive feedback from the learner's historical interaction records.

7.1 GAN and Micro Learning

According to the reviewed prior studies in Chapter 2, in addition to the combination of generation and discrimination, we can identify advantages of GAN that naturally comply with the demands of the development of recommender system for the micro learning service.

The experimental results of two pioneer works, CFGAN and IRGAN, have demonstrated the potentials of GAN when comparing to many state-of-the-art recommender systems. The representative statistics details are shown in Table 7.1 [Chae, et al., 2018] and Table 7.2 [Wang, et al., 2017]. In these two studies, top *N* results are evaluated separately for each model; ndcg@*N*, mrr@*N*, p@*N*, and r@*N* are the shorthand for nDCG, MRR, precision and recall scores for different *N* values. The experimental results from prior studies have shown that the GAN-based recommender system has better performance than various other state-of-the-art recommender systems in different application scenarios.

Nevertheless, to fill the gaps of the micro learning service, which have been discussed in Chapter 2, the feasible solutions towards the motivations raised in applying GAN-based recommendation strategy in micro learning can be summarised as follows:

Table 7.1 Item Recommendation Results on Movielens Dataset

Model	p@20	r@20	ndcg @10	mrr@20
ItemPop	0.138	0.251	0.195	0.292
BPR	0.236	0.287	0.380	0.574
FISM	0.285	0.353	0.429	0.685
CADE	0.287	0.353	0.425	0.674
GraphGAN	0.151	0.260	0.249	0.312
IRGAN	0.221	0.275	0.368	0.523
CFGAN	**0.294**	**0.360**	**0.433**	**0.693**

1. Imbalanced data distribution and cold-start problem frequently occur in recommendation scenario of micro learning service. Data augmentation is a straightforward way to supplement the system with extra information. As aforementioned in Chapter 2 (Section 2.5.4), GAN naturally has the advantages of generating lifelike data. GAN-based data augmentation models have been demonstrated that they can boost the quality of recommendation result, like using negative samples for precisely modelling user behaviour [Gao, et al., 2019] and new interactions for less active users [Wang, et al., 2019].
2. As discussed in Chapter 2 (Section 2.5.3), with proper adjustments in the loss function, GAN-based models can precisely rank the recommendation list. The ranking ability is significant for a recommender system when a learner is not familiar with the discipline he wishes to study, or ad hoc online resources.
3. Diversity is one prominent characteristic in the informal learning scenario. It is proven in the research of DP-GAN (discussed in Chapter 2 (Section 2.5.3), GAN is one possible solution to

Table 7.2 Item Recommendation Results on Netflix Dataset

Model	p@5	p@10	ndcg @5	ndcg@10	mrr@all
MLE	0.2945	0.2777	0.3011	0.2878	0.5058
BPR	0.2933	0.2774	0.2993	0.2866	0.5040
LambdaFM	0.3790	0.3489	0.3854	0.3624	0.5857
IRGAN	**0.4335**	**0.3923**	**0.4404**	**0.4097**	**0.6371**

balance the trade-off of relevance and diversity.

4. As mentioned in Chapter 2 (2.5.1), GAN has been used in various generation tasks from natural language to the video stream and demonstrate its outstanding creativity. Micro learning mainly targets the self-directed learner to quickly gain required knowledge under few constraints. In that sense, the learning style and learning resources could vary considerably in a different context. Complimented by such creativity, GAN also has the potential to positively influence the learning outcome of such informal learning.

7.2 The Pilot Experiment

7.2.1 Experimental Configurations

In order to find out whether the merit of the GAN-based framework can bridge the gaps in micro learning service, as we discussed before, we compare the effectiveness of two classic neural network-based recommender systems with or without using the adversarial learning strategy. The dataset used in this pilot experiment is collected from ZhiHu, the detail of this dataset please refer to the experiment section in Chapter 5.

Two classic neural network-based recommender systems are involved in the experiments, one is multi-layer perceptron (MLP) and another one is stacked autoencoder (SAE) [Tallapally, et al., 2018]. Both models can mine latent information from interaction records. In the experiments, these two models are trained with and without adversarial learning strategy. For the adversarial mode, both models are used as generators to select micro learning units from online resources that a user might be interested in.

7.2.2 Results

The pilot experiment results are shown in Table 7.3. The first and the third rows are the results produced by the models trained using the adversarial learning strategy, while the second and the fourth are trained without using the adversarial learning strategy. We can easily observe that using an adversarial learning strategy to train the model can further boost the model performance. Moreover, we can see that the simpler model MLP outperforms the complex model SAE without using the adversarial learning strategy. However, when using adversarial learning, SAE outperforms all the models to a great extent.

After analysis, we suggest that using the adversarial learning strategy can force the information flowing and exchanging between the generator and the discriminator, and such a process is helpful for the recommender system to avoid getting stuck in the local optimum point. Hence, we can conclude that using adversarial learning (i.e., GAN) has the potential to further boost the performance of a recommender system for micro learning service.

7.2.3 Research Gaps and the Application Background of a Novel GAN Model

As discussed in Chapter 5, a well-designed recommender system is a key factor to personalised online learning service. However, the recent mainstream studies in this field only focus on improving the performance of a recommender system by adding new features. For example, features about user/item's profile are used in [Lin, et al., 2020], time-series information is used in [Feng, et al., 2019], and DCN is based on interactions of various features [Wang, et al., 2017]. One drawback of this strategy is that sometimes some of the features are hard to be captured in some application scenarios or are not captured in many exist (public) datasets. Hence, in this research, based on the application background of online knowledge sharing service, we trace back to the fundamental issue about the recommender system of how to effectively utilise the user-resource interaction logs.

For a personalised online learning service, the recommended resources that match a learner's interest are regarded as positive results and those which do not are regarded as negative results. False-positive and false-negative are two different types of errors that affect the robustness of a recommender system. One of the key challenges for a recommendation task in online learning is to balance the false-negative rate (FNR) and the false-positive rate (FPR) of the recommendation

Table 7.3 Comparison of Models Using and not Using Adversarial Learning

Model	ndcg@10	mrr@10	p@10	r@10	ndcg@20	mrr@20	p@20	r@20
MLP (GAN)	**0.4116**	**0.6501**	**0.3431**	**0.2223**	**0.4092**	**0.6536**	**0.2764**	**0.3396**
MLP	0.407	0.6414	0.3417	0.2094	0.3934	0.6438	0.2712	0.304
SAE (GAN)	**0.435**	**0.6827**	**0.3595**	**0.2371**	**0.4269**	**0.6855**	**0.284**	**0.3497**
SAE	0.3941	0.6385	0.3261	0.2151	0.3934	0.6423	0.2638	0.326

results [Schröder, et al., 2011]. This usually requires two ‘opposite’ strategies to control and suppress these two errors: one for guiding the model to identify as many positive results as possible and the other for preventing the model from choosing negative results. Here are two well-known extreme situations: When the model regards all the candidate results as positive, all positive results are found, but the FPR is the highest. Conversely, when the model regards all the candidate results as negative, it avoids all negative results, but the FNR reaches the highest. For an effective recommender system, as many positive and as few negative results are the aim.

Moreover, in the recommendation scenario, there are massive implicit non-positive feedbacks, such as unrated items from a learner. The unrated items can be the negative learning materials that a learner is not interested in or the items that a learner has not yet interacted with. Properly utilising such non-positive information can more precisely profile a user’s learning preference. Therefore, balancing the trade-off between FNR and FPR and distilling the non-positive information can be the key to improve the recommendation effectiveness for the online learning services.

With the advantages of generalisation and flexibility, we hold that the GAN-based model has the potential to further exploring latent information from user-item interaction logs. As discussed in the prior studies [Lin, et al., 2019, Wu, et al., 2015] that the online learning scenario involves massive vague information, and for a knowledge-sharing platform it always covers a wide range of disciplines. A recommender system for online knowledge service is required to be flexible and generalise enough. Hence, in this research, we propose an adversarial learning framework to recommend personalised learning resource to online learners engaged with the online knowledge sharing service. The proposed framework is composed of three generators and one discriminator. With well-designed loss functions, this framework can suppress the FPR and FNR at the same time and also involve learner’s implicit non-positive feedbacks.

7.3 Model Design

The overview of our proposed model is introduced firstly in this section. Next, the loss functions used in this research is presented and explained. Last, the designing details of the generators and the discriminator are discussed.

7.3.1 The Framework Overview

In this study, a novel GAN framework is designed to solve the well-known collaborative filtering problem [Schafer, et al., 2007] for the micro learning service. The proposed framework consists of three generators and one discriminator. These generators have different goals and are optimised by different loss functions. The first generator G_1 aims to avoid selecting learning resources that do not meet the user's preference (minimise the FPR). The second generator G_2 aims to select as many as possible learning resources that meet the user's preference (minimise the FNR). The third generator G_3 aims to select the learning resources that a user will not be interested in. The generators are constructed by using the stacked autoencoder and the discriminator is constructed by using multi-layer perceptron conditioned by learners' historical learning activities. The network structures of the discriminator and the generator are shown in Figure 7.1.

The training workflow is shown in Figure 7.2, which is partially based on the prior work of cGAN [Mirza, et al., 2014]. X is the input information for the discriminator and the generators. The difference between the proposed model and cGAN is that our model has three generators G_1, G_2, and G_3. Therein, G_1 and G_2 generate (select) positive learning resources R_1 and R_2, while G_3 generates (selects) negative learning resources R_3 from the candidates. Formally, these processes can be represented as $G_1(X) \rightarrow R_1$, $G_2(X) \rightarrow R_2$, and $G_3(X) \rightarrow R_3$. During the adversarial training, the

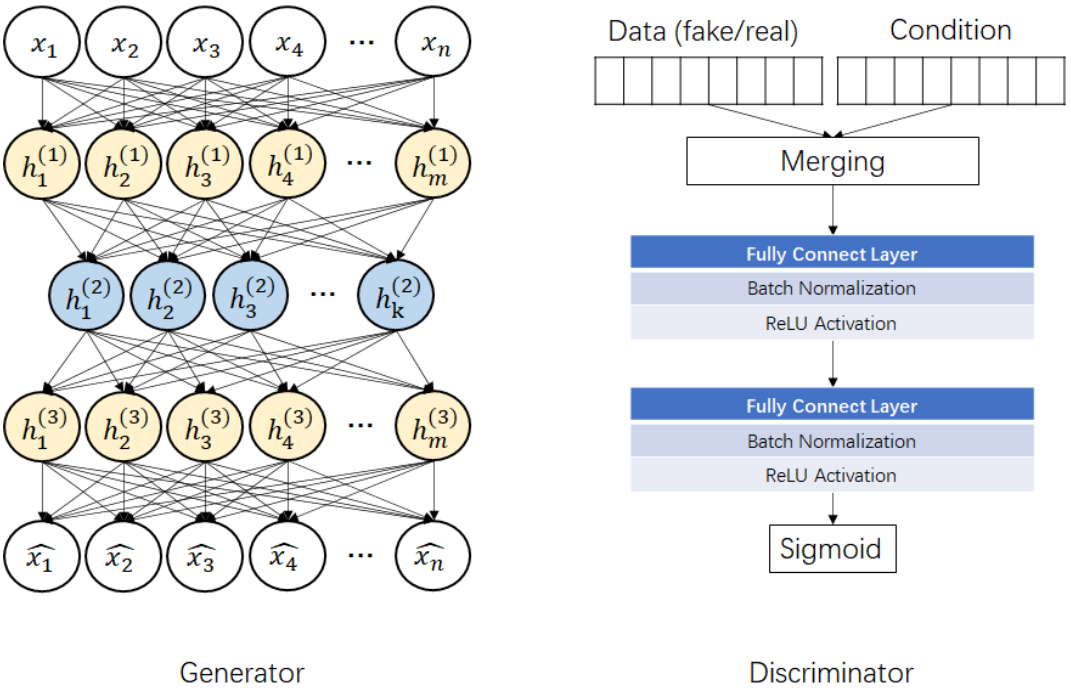

Figure 7.1 Network Structure of the Discriminator and the Generator

discriminator D is assigned to distinguish four types of selected learning resources R_0, R_1, R_2, and R_3, where R_0 denotes the ground truth learning resources that the user is interested in, and the R_1, R_2, and R_3 are the fake learning resources selected by different generators.

Specifically, the proposed three generators are jointly trained through the cGNA framework, which avoids the generators being optimised in different ways or selecting inferior learning resources to the target user. During the training step, the discriminator acts as a medium, connecting all the generators and allowing the distilled information to flow among them. Such type of information exchanging strategy can guarantee mining implicit, non-positive feedback and reducing FPR and FNR simultaneously.

After training, each of the generators[17] can be used as a recommender model to select learning resources that the user might be interested in. As all the generators have been trained jointly through the adversarial training process and the entire information flow is shared among them, each generator can recommend reasonable results consequently. However, in practice, we intend to use all three generators to recommend learning resources and fuse their outputs into the final recommendation results to guarantee better robustness. The recommendation (prediction) stage is shown in Figure 7.3.

7.3.2 Loss Functions for the Proposed Model

In order to better optimize our model, four loss functions $\mathcal{L}$ were carefully designed, the adversarial loss between the generators and a discriminator, the data loss for suppressing FPR and FNR, the

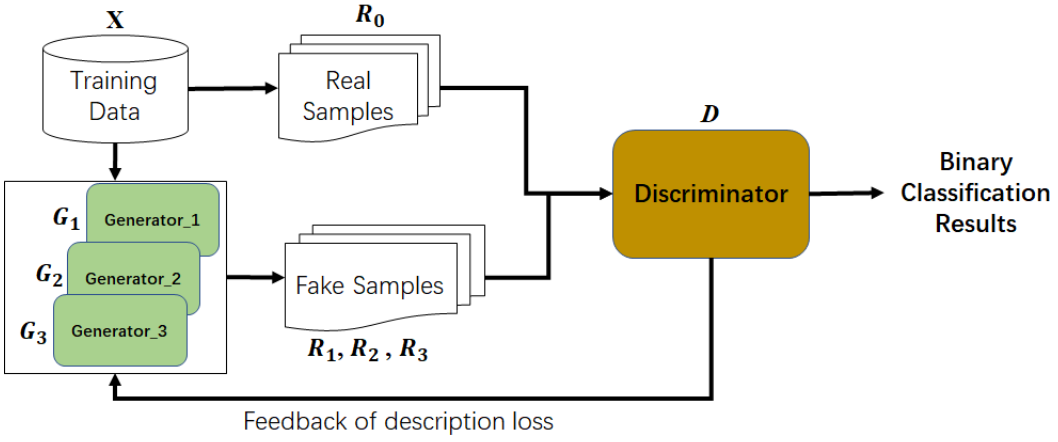

Figure 7.2 Network Structure of the Discriminator and the Generator

[17] For the third generator G_3, we need to choose the leaning materials with the lowest scores, because G_3 is trained to select negative samples.

consistency loss for forcing different generators to be optimized towards the same goal, and the non-positive feedback loss for modelling unrated resources. In summary, the complete objective of the proposed model is formulated as Equation (7.1) (G_1^*, G_2^*, G_3^*, and D^* represent for the generators and discriminator, four different types $\mathcal{L}$ are the loss functions used in this study):

$$(G_1^*, G_2^*, G_3^*, D^*) = \arg \min_{G_1,G_2,G_3} \max_{D} (\mathcal{L}_{cGAN} + \mathcal{L}_{Data} + \mathcal{L}_{consistency} + \mathcal{L}_{nf}) \quad (7.1)$$

7.3.2.1 Adversarial Loss

The adversarial loss models the competition manner between the generators and the discriminator. Different from the commonly used adversarial loss function like the one proposed in IRGAN one [Wang, et al., 2017], the adversarial loss used in this study contains four terms according to the usage of three generators and one discriminator, which can be formulated as Equation (7.2).

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x\sim p_{data}}[logD(c, X_0)] + \mathbb{E}_{\hat{x}\sim p_\phi}[\log(1 - D(c, G_1(x)))] + \mathbb{E}_{\hat{x}\sim p_\phi}[\log(1 - D(c, G_2(x)))] + \mathbb{E}_{\hat{x}\sim p_\phi}[\log(1 - D(c, G_3(x)))] \quad (7.2)$$

Where the first term corresponds to the discriminator and the last three terms correspond to the generators, c is the condition for cGAN, x is input information for the generators and X_0 is the ground truth. Maximizing this loss function guides the generators to generate more vivid results which are close to the ground truth. Minimizing this loss function enhances the discrimination ability of the discriminator to tell the difference between the generated data and the real one.

7.3.2.2 Data Loss

The second loss function is the data loss, which measures the FPR and FNR. The data loss for the first generator, G_1, and second generator, G_2, is formulated in Equation (7.3) and (7.4), respectively.

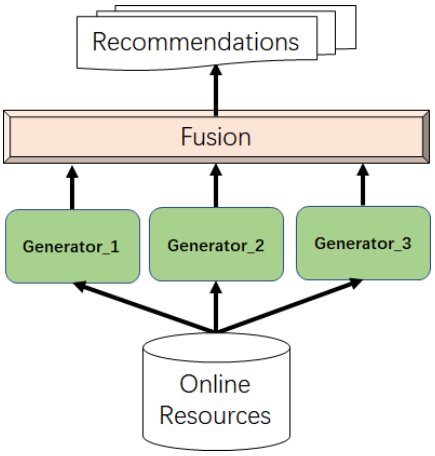

Figure 7.3 The Recommendation Workflow

The final data loss is the sum of these two losses, as Equation (7.5). It is worth noting that as the third generator G3 was designed for selecting negative samples, we did not calculate data loss for G_3.

$$\mathcal{L}_{Data_1}(G_1) = \frac{1}{n}\sum_{i=1}^{n}(\lambda_1 FN_{1i} + FP_{1i}) \tag{7.3}$$

$$\mathcal{L}_{Data_2}(G_2) = \frac{1}{n}\sum_{i=1}^{n}(FN_{2i} + \lambda_2 FP_{2i}) \tag{7.4}$$

$$\mathcal{L}_{Data} = \mathcal{L}_{Data_1} + \mathcal{L}_{Data_2} \tag{7.5}$$

Where λ_1 and λ_2 are weighting parameters which balance the influences of false-negative results and false-positive results, respectively. This weighting strategy makes two generators focus on distilling different types of information; where one pays more attention to selecting as many positive results as possible, and the other pays more attention to avoiding selecting negative results. One noteworthy point is that the data losses formulated above still focus on different objective FPR and FNR, respectively. The FPR regularized by a small FNR value (controlled by λ_1) in G_1 can achieve better initialization for training (similar for G_2).

7.3.2.3 Consistency Loss

Argued in the prior work of [Wang, et al., 2019], the generators are optimised towards the gradient, which has strong randomicity based on different initial settings. This could lead the generators to converge in different ways to cause discrepant results. To avoid this problem, a consistency loss is used to bind these generators and to further force the medium training information to flow among them. *L2* norm is used to measure the difference between the outputs of two generators. Thus, consistency loss for each generator is formulated as Equation (7.6), (7.7), and (7.8).

$$\mathcal{L}_{consistency}(G_1) = \alpha_1||G_1(x) - G_2(x)||_2^2 + \alpha_2||G_1(x) - G_3(x)||_2^2 \tag{7.6}$$

$$\mathcal{L}_{consistency}(G_2) = \alpha_3||G_2(x) - G_1(x)||_2^2 + \alpha_4||G_2(x) - G_3(x)||_2^2 \tag{7.7}$$

$$\mathcal{L}_{consistency}(G_3) = \alpha_5||G_3(x) - G_1(x)||_2^2 + \alpha_6||G_3(x) - G_2(x)||_2^2 \tag{7.8}$$

Where α_i is a hyper parameter that indicates the weight of each loss value.

7.3.2.4 Non-positive Feedback Loss

Inspired by the idea of zero-reconstruction used in a prior CFGAN [Chae, et al., 2018], in this study, we applied this idea to formulate the loss function for implicit non-positive feedback and applied it to each generator. For each user, a certain number of non-positive feedbacks were randomly selected. As discussed in section 7.3, using non-positive feedbacks can better reflect what a user does not

prefer. Equation (7.9) accumulates the loss of the prediction of non-positive feedbacks.

$$\mathcal{L}_{nf}(G) = \beta \sum_j (x_{uj}^- - \widehat{x_{uj}^-})^2 \tag{7.9}$$

Where x_{uj}^- represents a randomly selected sample which the user u has not given any positive feedbacks to the learning resource j, $\hat{x}_{uj}$ is the prediction made by the model, and β is a hyper parameter that controls the importance of this loss value.

7.3.3 The Generator and the Discriminator

It has been demonstrated in many existing representative studies that the autoencoder and its variants can precisely generate (select) the appropriate candidates based on a user's historical online behaviours. For example, the collaborative variational autoencoder (CVAE) model used in the study [Li, et al., 2017] can learn deep latent representations from content data in an unsupervised manner and capture the implicit relationships between users and items. A stacked denoising autoencoder is used in [Liu, et al., 2018] to learn user and item features from auxiliary information. In addition, an augmented variational autoencoder shows outstanding performance in dealing with auxiliary information and modelling the implicit user feedback in work [Lee, et al., 2017]. Hence, for simplicity, we design the generators directly follow the idea of using autoencoders. While the design of the generator is beyond the scope of this study, we only need to confirm that the used network can generate (select) reasonable learning resources based on a learner's historical learning activities. For experimental simplicity and mathematical convenience, the generators used in this study are constructed using stacked autoencoders (in Figure 7.1 left).

The discrimination process can be roughly regarded as a binary classification task. The design of the discriminator follows the idea proposed in the original work of cGAN [Mirza, et al., 2014], by considering not to make the discriminator so powerful that it dominates the adversarial training process and hinders the optimising of the generators. Therefore, a simple multi-layer fully connected perceptron with input condition is used as the discriminator in this study (as shown in Figure 7.1 right).

Theoretically, any conventional or state-of-the-art neural networks can be used in a GAN framework as long as their training process is in an adversarial manner. This offers huge flexibility to our model, which can be further optimised by simply replacing the network structure of each generator or discriminator based on the application requirement or the data source.

7.4 Experiment and Analysis

The experiment is conducted by using a real-world dataset to verify the effectiveness of the proposed model.

7.4.1 Dataset

The dataset collected from the ZhiHu Platform is still used in this part. To emphasise the fundamental issue of a recommender system that the recommendations are produced based on the historical interaction records between the user and resource, in this study, we only use the interaction records to construct a recommender system. Utilising other side information, such as answers to the questions and user's profile, is beyond the scope of this study and will be investigated in our future research. The selected dataset contains about 1.8 million different questions and users. The distribution of the (user, learning resource) interactions are extremely imbalanced, with a sparsity degree greater than 99%, which means that most users have only accessed a tiny minority of online learning resources. In addition, due to the difference of learning requirements and preferences, there is little overlap between different users' interaction histories with the online learning resources. Such sparsity and imbalance reflect the characteristics of personalised online learning services in the context of big data.

7.4.2 Baseline Comparison Models

For a fair comparison, we only include the models which only utilise the user-item interaction records. The models that make use of other types of features are excluded in the experiments. Firstly, to demonstrate the superior performance of the designed GAN framework, the proposed model is compared with two representative GAN-based recommendation strategies: IRGAN and CFGAN [Chae, et al., 2018, Wang, et al., 2017]. Moreover, to demonstrate the challenges brought by big data and personalisation of online learning service, the proposed model was also compared with several conventional recommender systems: Collaborative filtering, SVD++, Factorization Machine, and stacked autoencoder.

1. IRGAN. IRGAN is the pioneering work that was built upon the adversarial learning to handle the recommendation task., which showed the potential of applying the GAN framework to general information retrieval tasks, such as web search, item recommendation and question

answering [Wang, et al., 2017].

2. CFGAN. CFGAN is the enhanced version of IRGAN, which used a vector-wise training strategy. CFGAN demonstrates its outstanding performance in consistent and universal recommendation accuracy in comparison with the state-of-the-art recommenders [Chae, et al., 2018].
3. SVD++. As the extension of singular value decomposition (SVD), SVD++ takes into account implicit rating information [Koren, 2008].
4. Factorization Machine (FM). As a more general approach to model the interactions between users and items, FM shows good resistance in sparsity problem [Rendle, 2010]. A large number of state-of-the-art deep learning recommender system (such as AFM, deepFM, and DCN) are based on the idea of FM.
5. Stacked autoencoder (SAE). It is a deep neural network technique to mine latent information from both user and item sides [Tallapally, et al., 2018].
6. Collaborative filtering (CF). CF is the most classic recommendation model. In this study, the CF model is enhanced by involving z-score normalisation to better describe the degree of similarity between two users.

7.4.3 Evaluation Metrics

According to the prior studies and the pilot experiment, we select five representative evaluation metrics to evaluate the proposed model from different perspectives.

Recall, Precision and F1-score. Recall reflects the fraction of the total amount of positive learning materials which interested by the learner are selected by the model. Precision indicates the amount of positive learning materials in the selected learning materials. F1-score considers both Precision and Recall value, which is often used in the field of information retrieval.

Normalized Discounted Cumulative Gain (nDCG). nDCG considers the different importance in positions of the selected learning resources, which reflects the quality of the ranking results. This metric is formulated as Equation (7.10) and (7.11).

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{7.10}$$

$$IDCG_p = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i - 1}}{log\ (i+1)} \tag{7.11}$$

Mean Reciprocal Rank (MRR). MRR also reflects the ranking quality of the recommendation results.

It is formulated as the multiplicative inverse of the first correct result in Equation (7.12) below:

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank_i} \tag{7.12}$$

Here, $\mathrm{rank_i}$ is the ranking position of the first relevant learning resource for the i-th user.

In general, the metrics of Precision, Recall and F1-score provide indications about to what extent the correct results are selected by the model, while nDCG and MRR measure the model's ranking ability. Moreover, in this study, recall@N, precision@N, nDCG@N, and MRR@N were further used to evaluate the top N results selected by the model.

7.4.4 Implementation Settings

All the baselines and the proposed model are implemented using the PyTorch framework [Paszke, et al., 2019]. The number of hidden layers for the fully connected perceptron and the discriminator D is set to three; the number of hidden layers for SAE and the generator Gs is set to 4. Batch normalization is used with batch number 256 before the activation function at each layer during the training process. The output of each hidden layer is activated by the ReLU function, and the final result is activated by the sigmoid function. Adam optimizer is used for all neural networks with a learning rate of 0.0001. The parameters are initialized by using the Xavier strategy [Glorot, et al., 2010]. As four networks (three generators and one discriminator) are involved in the proposed model, each optimisation procedure could require different numbers of training epoch. Therefore, the early-stop strategy is used to prevent the overfitting of each model. All the baselines are constructed strictly following their original studies. Other settings use the default settings provides by the PyTorch framework.

7.4.5 Results and Discussions

The performances of each model are shown in Table 7.4 and Table 7.5 (P@k and R@k represent the Precision and Recall value for top k recommended results, respectively).

7.4.5.1 Model Comparisons

All metrics of the CF model are comparatively unsatisfactory, even if a z-score normalisation strategy is applied to better represent the difference between the users. This result reflects a typical case of the failure of solely utilising user-item interaction logs for recommendation purpose. Our analysis suggests that this failure is caused by the highly sparse high dimensional data. Such input

data led to the failure of similarity measurement [Houle, et al., 2010], which is the core part of the CF model.

Comparing with CF, SVD++ show a clear improvement, but it is still unsatisfying. Such improvement is made by adding the stage of dimension reduction during the modelling process. However, as discussed in [Rendle, 2010], without tuning hyperparameters carefully, SVD++ is not applicable to the general task and has special requirements for the input data. The generalisation ability of a recommender system is a key parameter for the informal learning service, as the type, format, and discipline of the learning resources might vary greatly. In prior research [Rendle, 2010] it is declared that FM could be a more general approach. Compared with SVD++, in our experiment, FM shows a big improvement in all metrics.

Moreover, figures in Table 7.4 and Table 7.5 show that the deep neural network-based models (SAE, IRGAN, CFGAN, and the proposed model) achieve remarkable improvement in comparison with the model with a shallow network (e.g. FM) or the ones without neural network (e.g. CF and SVD++). These results highlight the importance of using complex non-linear transformation to model the learner's preference in the context of an open learning environment. Another notable

Table 7.4 Experiment Results (Precision and Recall)

Metric / Model	P@5	P@10	P@15	P@20	R@5	R@10	R@15	R@20
CF (with Z-score)	0.0094	0.0078	0.0106	0.0126	0.0012	0.0021	0.0051	0.0077
SVD++	0.0730	0.0650	0.0636	0.0624	0.0137	0.0254	0.0326	0.0407
FM	0.1774	0.1408	0.1312	0.1213	0.0732	0.1185	0.1464	0.1673
SAE	0.3431	0.3154	0.2825	0.2564	0.1601	0.2043	0.2633	0.3114
IRGAN	0.3322	0.3037	0.2688	0.2452	0.1493	0.1911	0.2468	0.2946
CFGAN	0.3936	0.3546	0.3098	0.2796	0.1804	0.2252	0.2886	0.3362
Proposed Model	**0.4086**	**0.3701**	**0.3247**	**0.2923**	**0.1930**	**0.2392**	**0.3072**	**0.3572**

Table 7.5 Experiment Results (MRR and NDCG)

Metric / Model	MRR@5	MRR@10	MRR@15	MRR@20	ndcg @5	ndcg @10	ndcg@15	ndcg @20
CF (with Z-score)	0.0341	0.0374	0.0410	0.0437	0.0118	0.0098	0.0116	0.0131
SVD++	0.1790	0.1963	0.2000	0.2027	0.0780	0.0714	0.0703	0.0702
FM	0.3409	0.3519	0.3523	0.3662	0.1956	0.1870	0.1855	0.1833
SAE	0.6124	0.6173	0.6196	0.6207	0.3880	0.3791	0.3763	0.3779
IRGAN	0.5797	0.5847	0.5875	0.5902	0.3706	0.3608	0.3548	0.3539
CFGAN	0.6656	0.6737	0.6718	0.6737	0.4430	0.4275	0.4128	0.4181
Proposed Model	**0.6909**	**0.6931**	**0.6966**	**0.6975**	**0.4635**	**0.4488**	**0.4411**	**0.4403**

result is that the SAE outperforms the IRGAN. The generator in IRGAN is constructed by MLP. Hence, one possible reason could be the inferior non-linear modelling ability of perceptron compared with SAE.

Lastly, the proposed model outperforms the CFGAN in all metrics. SAE is used in both CFGAN and the proposed model to construct the generator. The main difference between CFGAN and our proposed model is the additional generators used in our model with carefully designed loss functions. Hence, we can infer that our model and the loss functions do further improve the model performance.

7.4.5.2 Effectiveness of each Generator

The optimisation process of the proposed model is shown in Figure 7.4. Due to the difficulty to directly compare Recall and Precision values, we use F1-score together with MRR and NDCG to measure our model performance during the training process. G_1, G_2, and G_3 are the three generators in the proposed GAN framework, respectively. To guarantee better robustness, the Gs is the model that combines all three generators' outputs.

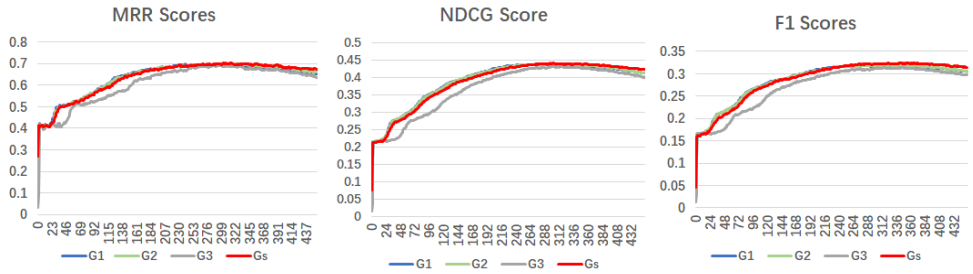

Figure 7.4 Performance of the Proposed Model During the Training Procedure

We can clearly see the converging optimisation trends from all the models (Figure 7.4), i.e., the generators all reach the optimal states at similar training epochs. The performance of G_3 is slightly worse in comparison with the others during the training process. However, when reaching the optimal states, all the generators show similar performances. These results verify the hypotheses proposed earlier that the proposed GAN framework and the associated loss functions can ensure that all information can be shared among the four generators. After training, each generator can make reasonable recommendations independently and together.

7.5 Summary

This chapter discusses the connections between the GAN technique and the recommendation task for the micro learning service. A pilot experiment based on the application of online knowledge sharing service has verified our assumption that the GAN-based model does have the potential to further improve the recommendation effectiveness of micro learning service.

Based on the specific application scenario, we further propose a novel GAN based recommendation framework. The model can fully explore the latent information from the simple user's historical interaction records with online learning resources with well-designed loss functions. By comparing with the state-of-the-art GAN-based recommender systems and conventional recommendation strategies, we have confirmed the robustness of the proposed model when facing the challenges of extreme data sparseness and unbalanced distribution in the existing platforms such as Quora and Zhihu.

In the future, we will continue investigating how to suit the GAN-based model to solve recommendation problem for micro learning service, considering other performance metrics, such as model training time, which has been considered still challenging in the research and development of

GAN based recommendation system.

Chapter 8 Conclusion and the Future Direction

This thesis is summarised in this chapter. Also, the future research directions are also suggested in this chapter.

8.1 Summary of Contributions in the Previous Chapters

In this thesis, we have introduced the research on realising the micro learning service through A.I. techniques. An intelligent micro learning system is proposed with the details of each intelligent component and the involved data flow. All the proposed models are evaluated with the-state-of-the-art solutions and real-world datasets.

In the first chapter, as the commence of this research, we have discussed the application background of the micro learning service. What is micro learning and why micro learning is significant to us are first introduced. This service aims to facilitate personalised online learning activities by using learner's fragmented spare time. Based on the prior work from the related areas, we conclude that as the product of the big data era, the micro learning service has to be deployed in the context of massive users and online resources. Then, we pin down the research challenges and problems of promoting this online learning service, which are the conflicts between handing massive online information and offering convenient real-time service. An A.I. embedded system can maximise the reduction of the labour force of managing massive online information and automatically provide the required service flexibly. To this end, the research objectives and contributions of this thesis are highlighted in Section 1.3.

A comprehensive literature review is made in the following Chapter 2. In this chapter, we firstly discuss the prior studies about the micro learning service, which include its definitions, characteristics, advantages, and the remained research gaps. It is safe to conclude that micro learning service is in line with the development trend of technologies and modern life, but it still requires much more work and research to be perfect. By breaking down the entire micro learning workflow into three modules, we then discuss the recent studies about the pre-processing of raw online resources, recommendation strategies, and dropout rate prediction methods. Technical backgrounds,

application scenarios, data requirements, and the comparison of different models are involved in the review. Recent research about the GAN technique is also reviewed in this chapter. The reviewed studies mainly cover the representative GAN milestones, the pioneering work of applying GAN to the recommendation task, and two branches of research about GAN-based recommender system. Based on the reviewed studies, we conclude that GAN has the potential to further boost the recommender system for the micro learning service.

The big picture of our proposed micro learning system is introduced in Chapter 3. In this chapter, the system framework is demonstrated from different perspectives which are the high-level point of view, the resource-side perspective, and the user-side perspective. After that, the involved data sources of the proposed system are discussed from their types, characteristics, and utilities. We also analyse the isolated data problem in this research area.

In Chapter 4, we have proposed two novel models for the pre-processing module of the micro learning system. The first model is for extracting valuable information from a sequence of the data stream. According to the experiment results, we conclude that our model outperforms the mainstream solutions in both efficiency and robustness. We also highlight the significance of the fusion block and the CRF layer, and the usefulness of using the CNN layer to mine supplementary information. The second model proposed in this chapter is a CNN-based text analysis solution. This solution aims to use a general model which has the potential to interpret both text and visual information. The proposed model is constructed by utilising the mainstream CV models. The experiment results indicate that our proposed model shows the competitive performance when interpreting the short informal text compared to other mainstream NLP models.

A deep cross attention network is proposed for delivering personalised learning materials in Chapter 5. The proposed model combines different deep learning techniques, such as attention mechanism, cross-network, and residual connection. Such network architecture makes the model has the ability to automatically generate high-order feature interactions and distinguish the importance differences among the massive features. In the experiment, the proposed model is compared with AutoInt, FM, DeepFM, DCN, and AFM. All the baselines are the representative recommender systems and have been widely used in both industry and research fields. The experiment results that our model outperforms the above baselines in both effectiveness and efficiency.

For assessing the effectiveness of the student's online learning activity, a dropout rate prediction

model is proposed in Chapter 6. The proposed model predicts whether a student will drop out from an enrolled course based on his/her historical learning records of this course. This model consists of two different components (micro component and macro component), which separately model the different granularity of information. According to the conducted experiment, except demonstrating our model can outperform the selected baselines, we have verified the existence of the temporal pattern of the learning interaction records and proved the effectiveness of each proposed mechanisms. We also have analysed and discussed the characteristics of the temporal pattern from the interaction records, which are different to the ones from the areas of NLP and computational genomics. At the end of this chapter, we also give out some suggestions on the implication of real-world applications of two training methods of the proposed model.

The GAN technique for boosting the performance of the recommender system in the micro learning service is discussed in Chapter 7. The advantages of applying GAN optimisation strategy to a recommender system for micro learning service have between firstly analysed in this chapter, followed by a pilot study which proves the correctness of the above theoretical analysis. Then, we dive deep to design a novel GAN framework for optimising the recommender system. The proposed GAN framework consists of three generators and one discriminator, which ensure balancing the trade-off between FNR and FPR and distilling the non-positive information. In the meantime, to make sure the generators and the discriminator can focus on their own tasks, several loss functions are designed to guide the optimisation process. In the experiment, the proposed GAN-based solution is compared with two representative GAN-based recommender systems (CFGAN and IRGAN) and several mainstream solutions. According to the results, we have confirmed the robustness of the proposed model when facing the challenges of extreme data sparseness and unbalanced distribution.

8.2 Recommendation for the Future Research

For the direction of future research in the related area, we give out the following recommendations and suggestions:

As most models involved in the proposed system are data-driven, the idea behind these optimisation and analysis strategies significantly overlaps with the other data-driven research topics in the technology-enhanced learning (TEL) domain. Even though this thesis is under the topic of micro learning, many points derived from the discussion and analysis of this can be extended to other e-

learning related research topics. The data challenges mentioned in Section 2.3.3 and Section 3.2.3 also impede the development of the relevant research in other TEL fields. Hence, it is worthwhile to construct a complete public dataset. The construction process requires more efforts from researchers and institutions worldwide.

A micro learning system can be regarded as a big online ecosystem, which can cover the entire lifecycle of the online learning resources and multiple types of learning activities. This system is made up of many different intelligent models. Each model aims to solve a specific task. In this thesis, we only realise a portion of these intelligent models. Hence, it is worthwhile to continue developing other models for different tasks for future research, such as final grade prediction, learning path generation, and learner's knowledge level assessment.

For the model proposed in Section 4.2, we only focus on using the CV method to understand the informal short text. In the future, firstly, it is worth continuing to investigate the effectiveness of the proposed solution for long text understanding (like the text content of a course). To better design such a generic model for processing multimodal information, it also needs to investigate applying the CNN-based solution to solve other forms of information such as audio signals. In the meantime, researchers can also try to use NLP solutions reversely to solve other forms of information.

In Chapter 5, our proposed recommender system is a static model, the model structure and all the involved parameters are fixed once the training procedure is finished. However, the data stream from the micro learning platform can be dynamic. The learning requirement of a user might drift from time to time after he/she finishes some courses, or obtain a certain degree, or hands on a new task. And the popularity of the courses will also change with the development of each discipline. When dealing with the dynamic data, we need to update (retrain) the constructed static model from time to time based on the evolution rate and degree of the data. Hence, for future research, it is worthwhile to continue investigating how to make the model self-adjustable. A self-adjustable model should be able to identify the changes in the data and make necessary adjustments to itself automatically.

In Chapter 7, we have traced back to the fundamental issue of the recommender system about utilising the interaction records between the users and the online resources. A novel GAN-based optimisation strategy is proposed, which solely uses the interaction records to make recommendations. However, for many real-world scenarios, we can obtain many other types of supplementary information or side information from users and items, such as the user's gender and

occupation. Hence, it is necessary for future research to further investigate how to utilise such extra information to further enhance the recommendation results.

Bibliography or List of References

[Chae, et al., 2018] Dong-Kyu Chae, Jin-Soo Kang, Sang-Wook Kim, and Jung-Tae Lee: 'Cfgan: A generic collaborative filtering framework based on generative adversarial networks'. In. Proceedings of the 27th ACM international conference on information and knowledge management pp. 137-146, 2018

[Yang, et al., 2018] Deqing Yang, Zikai Guo, Ziyi Wang, Juyang Jiang, Yanghua Xiao, and Wei Wang: 'A Knowledge-Enhanced Deep Recommendation Framework Incorporating GAN-based Models'. In. 2018 IEEE International Conference on Data Mining (ICDM) pp. 1368-1373, 2018

[Bharadhwaj, et al., 2018] Homanga Bharadhwaj, Homin Park, and Brian Y Lim: 'Recgan: recurrent generative adversarial networks for recommendation systems'. In. Proceedings of the 12th ACM Conference on Recommender Systems pp. 372-376, 2018

[Zhou, et al., 2019] Fan Zhou, Ruiyang Yin, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Jin Wu: 'Adversarial point-of-interest recommendation'. In. The World Wide Web Conference pp. 3462-34618, 2019

[Wang, et al., 2019] Yang Wang, Hai-Tao Zheng, Wang Chen, and Rui Zhang: 'LambdaGAN: Generative Adversarial Nets for Recommendation Task with Lambda Strategy'. In. 2019 International Joint Conference on Neural Networks (IJCNN) pp. 1-8, 2019

[Wu, et al., 2019] Qiong Wu, Yong Liu, Chunyan Miao, Binqiang Zhao, Yin Zhao, and Lu Guan: 'PD-GAN: adversarial learning for personalized diversity-promoting recommendation'. In. Proceedings of the 28th International Joint Conference on Artificial Intelligence pp. 3870-3876, 2019

[Sun, et al., 2018] Geng Sun, Tingru Cui, Jianming Yong, Jun Shen, and Shiping Chen: 'MLaaS: a cloud-based system for delivering adaptive micro learning in mobile MOOC learning', IEEE Transactions on Services Computing, 2018, 11, (2), pp. 292-305

[Guo, et al., 2014] Philip J Guo, Juho Kim, and Rob Rubin: 'How video production affects student engagement: an empirical study of MOOC videos'. In. Proceedings of the first ACM conference on Learning@ scale conference pp. 41-50, 2014

[Anderson, et al., 2014] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec: 'Engaging with massive online courses'. In. Proceedings of the 23rd international conference on World wide web pp. 687-698, 2014

[Syeda-Mahmood, et al., 2001] Tanveer Syeda-Mahmood, and Dulce Ponceleon: 'Learning video browsing behavior and its application in the generation of video previews'. In. Proceedings of the ninth ACM international conference on Multimedia pp. 119-128, 2001

[Lin, et al., 2019] Jiayin Lin, Geng Sun, Tingru Cui, Jun Shen, Dongming Xu, Ghassan Beydoun, Ping Yu, David Pritchard, Li Li, and Shiping Chen: 'From ideal to reality: segmentation, annotation, and recommendation, the vital trajectory of intelligent micro learning', World Wide Web, 2019, 23, (3), pp. 1-21

[Kovachev, et al., 2011] Dejan Kovachev, Yiwei Cao, Ralf Klamma, and Matthias Jarke: 'Learn-as-you-go: new ways of cloud-based micro-learning for the mobile web'. In. International conference on web-based learning pp. 51-61, 2011

[Bruck, et al., 2012] Peter A Bruck, Luvai Motiwalla, and Florian Foerster: 'Mobile Learning with Micro-content: A Framework and Evaluation', Bled eConference, 2012, 25, pp. 527-543

[Mohammed, et al., 2018] Gona Sirwan Mohammed, Karzan Wakil, and Sarkhell Sirwan Nawroly: 'The effectiveness of microlearning to improve students' learning ability', International Journal of Educational

Research Review, 2018, 3, (3), pp. 32-38

[Bolettieri, et al., 2007] Paolo Bolettieri, Fabrizio Falchi, Claudio Gennaro, and Fausto Rabitti: 'Automatic metadata extraction and indexing for reusing e-learning multimedia objects'. In. Workshop on multimedia information retrieval on The many faces of multimedia semantics pp. 21-28, 2007

[Al-Shamri, et al., 2008] Mohammad Yahya H Al-Shamri, and Kamal K Bharadwaj: 'Fuzzy-genetic approach to recommender systems based on a novel hybrid user model', Expert systems with applications, 2008, 35, (3), pp. 1386-1399

[Wasid, et al., 2015] Mohammed Wasid, and Vibhor Kant: 'A particle swarm approach to collaborative filtering based recommender systems through fuzzy features', Procedia Computer Science, 2015, 54, pp. 440-448

[Fong, et al., 2008] Simon Fong, Yvonne Ho, and Yang Hang: 'Using genetic algorithm for hybrid modes of collaborative filtering in online recommenders'. In. Eighth International Conference on Hybrid Intelligent Systems pp. 174-179, 2008

[Bobadilla, et al., 2011] Jesus Bobadilla, Fernando Ortega, Antonio Hernando, and Javier Alcalá: 'Improving collaborative filtering recommender system results and performance using genetic algorithms', Knowledge-based systems, 2011, 24, (8), pp. 1310-1316

[Ujjin, et al., 2003] Supiya Ujjin, and Peter J Bentley: 'Particle swarm optimization recommender system'. In. Swarm Intelligence Symposium, 2003. SIS'03. Proceedings of the 2003 IEEE pp. 124-131, 2003

[Ujjin, et al., 2002] Supiya Ujjin, and Peter J Bentley: 'Learning user preferences using evolution'. In. Proceedings of the 4th Asia-Pacific conference on simulated evolution and learning, Singapore pp. 2002

[Acun, et al., 2021] Bilge Acun, Matthew Murphy, Xiaodong Wang, Jade Nie, Carole-Jean Wu, and Kim Hazelwood: 'Understanding training efficiency of deep learning recommendation models at scale'. In. 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA) pp. 802-814, 2021

[Rungsuptaweekoon, et al., 2017] Kanokwan Rungsuptaweekoon, Vasaka Visoottiviseth, and Ryousei Takano: 'Evaluating the power efficiency of deep learning inference on embedded GPU systems'. In. 2017 2nd International Conference on Information Technology (INCIT) pp. 1-5, 2017

[Eshach, 2007] Haim Eshach: 'Bridging in-school and out-of-school learning: Formal, non-formal, and informal education', Journal of science education and technology, 2007, 16, (2), pp. 171-190

[Lin, et al., 2019] Jiayin Lin, Geng Sun, Jun Shen, Tingru Cui, Ping Yu, Dongming Xu, Li Li, and Ghassan Beydoun: 'Towards the readiness of learning analytics data for micro learning'. In. International Conference on Services Computing pp. 66-76, 2019

[Sun, et al., 2018] Geng Sun, Tingru Cui, Jianming Yong, Jun Shen, and Shiping Chen: 'MLaaS: A Cloud-Based System for Delivering Adaptive Micro Learning in Mobile MOOC Learning', IEEE Transactions on Services Computing, 2018, (2), pp. 292-305

[Sun, et al., 2018] Geng Sun, Tingru Cui, Fang Dong, Dongming Xu, Jun Shen, Shiping Chen, and Jiayin Lin: '(WIP) Evaluation of a Cloud-Based System for Delivering Adaptive Micro Open Education Resource to Fresh Learners'. In. 2018 IEEE 11th International Conference on Cloud Computing (CLOUD) pp. 586-589, 2018

[Sun, et al., 2017] Geng Sun, Tingru Cui, Ghassan Beydoun, Shiping Chen, Fang Dong, Dongming Xu, and Jun Shen: 'Towards massive data and sparse data in adaptive micro open educational resource recommendation: a study on semantic knowledge base construction and cold start problem', Sustainability, 2017, 9, (6), pp. 898

[Guo, et al., 2017] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He: 'DeepFM: a factorization-machine based neural network for CTR prediction'. In. Proceedings of the 26th International Joint Conference on Artificial Intelligence pp. 1725-1731, 2017

[Ma, et al., 2016] Xuezhe Ma, and Eduard Hovy: 'End-to-end sequence labeling via bi-directional lstm-cnns-crf', arXiv preprint arXiv:1603.01354, 2016

[McCallum, et al., 2000] Andrew McCallum, Dayne Freitag, and Fernando CN Pereira: 'Maximum Entropy Markov Models for Information Extraction and Segmentation'. In. Icml pp. 591-598, 2000

[Zhu, et al., 2005] Jun Zhu, Zaiqing Nie, Ji-Rong Wen, Bo Zhang, and Wei-Ying Ma: '2d conditional random fields for web information extraction'. In. Proceedings of the 22nd international conference on Machine learning pp. 1044-1051, 2005

[Li, et al., 2018] Tao Li, Minsoo Choi, Kaiming Fu, and Lei Lin: 'Music sequence prediction with mixture hidden markov models', arXiv preprint arXiv:1809.00842, 2018

[Sutskever, et al., 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le: 'Sequence to sequence learning with neural networks'. In. Advances in neural information processing systems pp. 3104-3112, 2014

[Dyer, et al., 2015] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith: 'Transition-based dependency parsing with stack long short-term memory', arXiv preprint arXiv:1505.08075, 2015

[Huang, et al., 2015] Zhiheng Huang, Wei Xu, and Kai Yu: 'Bidirectional LSTM-CRF models for sequence tagging', arXiv preprint arXiv:1508.01991, 2015

[Yang, et al., 2018] Jie Yang, Shuailong Liang, and Yue Zhang: 'Design Challenges and Misconceptions in Neural Sequence Labeling'. In. Proceedings of the 27th International Conference on Computational Linguistics pp. 3879-3889, 2018

[Chen, et al., 2019] Jiahao Chen, Hang Li, Wenxin Wang, Wenbiao Ding, Gale Yan Huang, and Zitao Liu: 'A multimodal alerting system for online class quality assurance'. In. International Conference on Artificial Intelligence in Education pp. 381-385, 2019

[Lynch, et al., 2016] Corey Lynch, Kamelia Aryafar, and Josh Attenberg: 'Images don't lie: Transferring deep visual semantic features to large-scale multimodal learning to rank'. In. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp. 541-548, 2016

[Zhao, et al., 2016] Baoquan Zhao, Songhua Xu, Shujin Lin, Xiaonan Luo, and Lian Duan: 'A new visual navigation system for exploring biomedical Open Educational Resource (OER) videos', Journal of the American Medical Informatics Association, 2016, 23, (e1), pp. e34-e41

[O'Mahony, et al., 2019] Niall O'Mahony, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan, and Joseph Walsh: 'Deep learning vs. traditional computer vision'. In. Science and Information Conference pp. 128-144, 2019

[Krizhevsky, et al., 2017] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton: 'Imagenet classification with deep convolutional neural networks', Communications of the ACM, 2017, 60, (6), pp. 84-90

[Girshick, et al., 2015] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik: 'Region-based convolutional networks for accurate object detection and segmentation', IEEE transactions on pattern analysis and machine intelligence, 2015, 38, (1), pp. 142-158

[Girshick, 2015] Ross Girshick: 'Fast r-cnn'. In. Proceedings of the IEEE international conference on computer vision pp. 1440-1448, 2015

[Lu, et al., 2019] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan: 'Grid r-cnn'. In. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 7363-7372, 2019

[Bertinetto, et al., 2016] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr: 'Fully-convolutional siamese networks for object tracking'. In. European conference on computer vision pp. 850-865, 2016

[Simonyan, et al., 2014] Karen Simonyan, and Andrew Zisserman: 'Very deep convolutional networks for large-scale image recognition', arXiv preprint arXiv:1409.1556, 2014

[Bai, et al., 2020] Shuai Bai, Zhiqun He, Yuan Dong, and Hongliang Bai: 'Multi-hierarchical independent correlation filters for visual tracking'. In. 2020 IEEE International Conference on Multimedia and Expo (ICME) pp. 1-6, 2020

[He, et al., 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun: 'Deep residual learning for image recognition'. In. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 770-778, 2016

[Zhai, et al., 2018] Zenan Zhai, Dat Quoc Nguyen, and Karin Verspoor: 'Comparing CNN and LSTM character-level embeddings in BiLSTM-CRF models for chemical and disease named entity recognition', arXiv preprint arXiv:1808.08450, 2018

[Pawade, et al., 2018] D Pawade, A Sakhapara, M Jain, N Jain, and K Gada: 'Story scrambler-automatic text generation using word level rnn-lstm', International Journal of Information Technology and Computer Science (IJITCS), 2018, 10, (6), pp. 44-53

[Kim, 2014] Yoon Kim: 'Convolutional neural networks for sentence classification', arXiv preprint arXiv:1408.5882, 2014

[Wang, et al., 2016] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu: 'Cnn-rnn: A unified framework for multi-label image classification'. In. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 2285-2294, 2016

[Guo, et al., 2018] Yanming Guo, Yu Liu, Erwin M Bakker, Yuanhao Guo, and Michael S Lew: 'CNN-RNN: A large-scale hierarchical image classification framework', Multimedia Tools and Applications, 2018, 77, (8), pp. 10251-10271

[Yao, et al., 2017] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei: 'Incorporating copying mechanism in image captioning for learning novel objects'. In. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 6580-6588, 2017

[Wu, et al., 2015] Dianshuang Wu, Jie Lu, and Guangquan Zhang: 'A fuzzy tree matching-based personalized e-learning recommender system', IEEE Transactions on Fuzzy Systems, 2015, 23, (6), pp. 2412-2426

[Sikka, et al., 2012] Reema Sikka, Amita Dhankhar, and Chaavi Rana: 'A survey paper on e-learning recommender system', International Journal of Computer Applications, 2012, 47, (9), pp. 27-30

[Al-Hmouz, et al., 2012] Ahmed Al-Hmouz, Jun Shen, Rami Al-Hmouz, and Jun Yan: 'Modeling and simulation of an adaptive neuro-fuzzy inference system (ANFIS) for mobile learning', IEEE Transactions on Learning Technologies, 2012, 5, (3), pp. 226-237

[Zhao, et al., 2016] Qin Zhao, Yueqin Zhang, and Jian Chen: 'An improved ant colony optimization algorithm for recommendation of micro-learning path'. In. Computer and Information Technology (CIT), 2016 IEEE International Conference on pp. 190-196, 2016

[Chen, et al., 2017] Mengyuan Chen, Mingwen Tong, Chunmiao Liu, Meimei Han, and Ying Xia:

‘Recommendation of learning path using an improved ACO based on novel coordinate system’. In. Advanced Applied Informatics (IIAI-AAI), 2017 6th IIAI International Congress on pp. 747-753, 2017

[Fenza, et al., 2017] Giuseppe Fenza, Francesco Orciuoli, and Demetrios G Sampson: ‘Building Adaptive Tutoring Model Using Artificial Neural Networks and Reinforcement Learning’. In. Advanced Learning Technologies (ICALT), 2017 IEEE 17th International Conference on pp. 460-462, 2017

[Vygotsky, 1980] Lev Semenovich Vygotsky: ‘Mind in society: The development of higher psychological processes’ (Harvard university press, 1980. 1980)

[Murray, et al., 2002] Tom Murray, and Ivon Arroyo: ‘Toward measuring and maintaining the zone of proximal development in adaptive instructional systems’. In. International Conference on Intelligent Tutoring Systems pp. 749-758, 2002

[Sutton, et al., 1998] Richard S Sutton, and Andrew G Barto: ‘Reinforcement learning: An introduction’ (MIT press, 1998. 1998)

[Hoic-Bozic, et al., 2015] Natasa Hoic-Bozic, Martina Holenko Dlab, and Vedran Mornar: ‘Recommender system and web 2.0 tools to enhance a blended learning model’, IEEE Transactions on education, 2015, 59, (1), pp. 39-44

[Chen, et al., 2014] Wei Chen, Zhendong Niu, Xiangyu Zhao, and Yi Li: ‘A hybrid recommendation algorithm adapted in e-learning environments’, World Wide Web, 2014, 17, (2), pp. 271-284

[Rusak, 2017] Zoltan Rusak: ‘Exploitation of micro-learning for generating personalized learning paths’. In. DS 87-9 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 9: Design Education, Vancouver, Canada, 21-25.08. 2017 pp. 129-138, 2017

[Zhao, et al., 2016] Qin Zhao, Yueqin Zhang, and Jian Chen: ‘An improved ant colony optimization algorithm for recommendation of micro-learning path’. In. 2016 IEEE International Conference on Computer and Information Technology (CIT) pp. 190-196, 2016

[Sun, et al., 2017] Geng Sun, Tingru Cui, Jun Shen, Dongming Xu, Ghassan Beydoun, and Shiping Chen: ‘Ontological learner profile identification for cold start problem in micro learning resources delivery’. In. 2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT) pp. 16-20, 2017

[Sun, et al., 2018] Geng Sun, Tingru Cui, Dongming Xu, Jun Shen, and Shiping Chen: ‘A Heuristic Approach for New-Item Cold Start Problem in Recommendation of Micro Open Education Resources’. Proc. International Conference on Intelligent Tutoring Systems2018 pp. Pages

[Shu, et al., 2018] Jiangbo Shu, Xiaoxuan Shen, Hai Liu, Baolin Yi, and Zhaoli Zhang: ‘A content-based recommendation algorithm for learning resources’, Multimedia Systems, 2018, 24, (2), pp. 163-173

[Niemann, et al., 2013] Katja Niemann, and Martin Wolpers: ‘Usage context-boosted filtering for recommender systems in TEL’. In. European Conference on Technology Enhanced Learning pp. 246-259, 2013

[Hajri, et al., 2017] Hiba Hajri, Yolaine Bourda, and Fabrice Popineau: ‘MORS: A System for Recommending OERs in a MOOC’. In. Advanced Learning Technologies (ICALT), 2017 IEEE 17th International Conference on pp. 50-52, 2017

[Rendle, 2010] Steffen Rendle: ‘Factorization machines’. In. 2010 IEEE International Conference on Data Mining pp. 995-1000, 2010

[Fischer, et al., 2018] Thomas Fischer, and Christopher Krauss: ‘Deep learning with long short-term memory networks for financial market predictions’, European Journal of Operational Research, 2018, 270, (2), pp. 654-669

[Liu, et al., 2017] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang: 'Deep learning for extreme multi-label text classification'. In. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval pp. 115-124, 2017

[Cheng, et al., 2016] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, and Mustafa Ispir: 'Wide & deep learning for recommender systems'. In. Proceedings of the 1st workshop on deep learning for recommender systems pp. 7-10, 2016

[Guo, et al., 2017] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He: 'DeepFM: a factorization-machine based neural network for CTR prediction'. In. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence pp. 1725-1731, 2017

[Lian, et al., 2018] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun: 'xdeepfm: Combining explicit and implicit feature interactions for recommender systems'. In. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining pp. 1754-1763, 2018

[Wang, et al., 2017] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang: 'Deep & cross network for ad click predictions'. In. Proceedings of the ADKDD'17 pp. 12, 2017

[Song, et al., 2019] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang: 'Autoint: Automatic feature interaction learning via self-attentive neural networks'. In. Proceedings of the 28th ACM International Conference on Information and Knowledge Management pp. 1161-1170, 2019

[Xiao, et al., 2017] Jun Xiao, Hao Ye, Xiangnan He, Hanwang Zhang, Fei Wu, and Tat-Seng Chua: 'Attentional factorization machines: Learning the weight of feature interactions via attention networks'. In. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence pp. 3119-3125, 2017

[Vaswani, et al., 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin: 'Attention is all you need'. In. Advances in neural information processing systems pp. 5998-6008, 2017

[Hu, et al., 2018] Jie Hu, Li Shen, and Gang Sun: 'Squeeze-and-excitation networks'. In. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 7132-7141, 2018

[Huang, et al., 2019] Tongwen Huang, Zhiqi Zhang, and Junlin Zhang: 'FiBiNET: Combining Feature Importance and Bilinear feature Interaction for Click-Through Rate Prediction'. In. Proceedings of the 13th ACM Conference on Recommender Systems pp. 169-177, 2019

[Shan, et al., 2016] Ying Shan, T Ryan Hoens, Jian Jiao, Haijing Wang, Dong Yu, and JC Mao: 'Deep crossing: Web-scale modeling without manually crafted combinatorial features'. In. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining pp. 255-262, 2016

[Ziegler, et al., 2005] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen: 'Improving recommendation lists through topic diversification'. In. Proceedings of the 14th international conference on World Wide Web pp. 22-32, 2005

[Dwivedi, et al., 2018] Pragya Dwivedi, Vibhor Kant, and Kamal K Bharadwaj: 'Learning path recommendation based on modified variable length genetic algorithm', Education and Information Technologies, 2018, 23, (2), pp. 819-836

[Zhou, et al., 2018] Yuwen Zhou, Changqin Huang, Qintai Hu, Jia Zhu, and Yong Tang: 'Personalized learning full-path recommendation model based on LSTM neural networks', Information Sciences, 2018, 444, pp. 135-152

[Harper, et al., 2016] F Maxwell Harper, and Joseph A Konstan: 'The movielens datasets: History and context', Acm transactions on interactive intelligent systems (tiis), 2016, 5, (4), pp. 19

[Dorça, et al., 2017] Fabiano A Dorça, Vitor C Carvalho, Miller M Mendes, Rafael D Araújo, Hiran N Ferreira, and Renan G Cattelan: 'An Approach for Automatic and Dynamic Analysis of Learning Objects Repositories Through Ontologies and Data Mining Techniques for Supporting Personalized Recommendation of Content in Adaptive and Intelligent Educational Systems'. In. 2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT) pp. 514-516, 2017

[Kloft, et al., 2014] Marius Kloft, Felix Stiehler, Zhilin Zheng, and Niels Pinkwart: 'Predicting MOOC dropout over weeks using machine learning methods'. In. Proceedings of the EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs pp. 60-65, 2014

[Whitehill, et al., 2015] Jacob Whitehill, Joseph Williams, Glenn Lopez, Cody Coleman, and Justin Reich: 'Beyond prediction: First steps toward automatic intervention in MOOC student stopout', Available at SSRN 2611750, 2015

[Tang, et al., 2018] Cui Tang, Yuanxin Ouyang, Wenge Rong, Jingshuai Zhang, and Zhang Xiong: 'Time series model for predicting dropout in massive open online courses'. In. International Conference on Artificial Intelligence in Education pp. 353-357, 2018

[Wang, et al., 2017] Wei Wang, Han Yu, and Chunyan Miao: 'Deep model for dropout prediction in MOOCs'. In. Proceedings of the 2nd International Conference on Crowd Science and Engineering pp. 26-32, 2017

[Dalipi, et al., 2018] Fisnik Dalipi, Ali Shariq Imran, and Zenun Kastrati: 'MOOC dropout prediction using machine learning techniques: Review and research challenges'. In. 2018 IEEE Global Engineering Education Conference (EDUCON) pp. 1007-1014, 2018

[He, et al., 2015] Jiazhen He, James Bailey, Benjamin IP Rubinstein, and Rui Zhang: 'Identifying at-risk students in massive open online courses'. In. Twenty-Ninth AAAI Conference on Artificial Intelligence pp. 2015

[Amnueypornsakul, et al., 2014] Bussaba Amnueypornsakul, Suma Bhat, and Phakpoom Chinprutthiwong: 'Predicting attrition along the way: The UIUC model'. In. Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs pp. 55-59, 2014

[Al-Shabandar, et al., 2017] Raghad Al-Shabandar, Abir Hussain, Andy Laws, Robert Keight, Janet Lunn, and Naeem Radi: 'Machine learning approaches to predict learning outcomes in Massive open online courses'. In. 2017 International Joint Conference on Neural Networks (IJCNN) pp. 713-720, 2017

[Al-Shabandar, et al., 2017] Raghad Al-Shabandar, Abir Hussain, Andy Laws, Robert Keight, and Janet Lunn: 'Towards the differentiation of initial and final retention in massive open online courses'. In. International Conference on Intelligent Computing pp. 26-36, 2017

[Imran, et al., 2019] Ali Shariq Imran, Fisnik Dalipi, and Zenun Kastrati: 'Predicting Student Dropout in a MOOC: An Evaluation of a Deep Neural Network Model'. In. Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence pp. 190-195, 2019

[Feng, et al., 2019] Wenzheng Feng, Jie Tang, and Tracy Xiao Liu: 'Understanding dropouts in MOOCs'. In. Proceedings of the AAAI Conference on Artificial Intelligence pp. 517-524, 2019

[Koller, et al., 2017] Oscar Koller, Sepehr Zargaran, and Hermann Ney: 'Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs'. In. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 4297-4305, 2017

[Zhao, et al., 2017] Rui Zhao, Dongzhe Wang, Ruqiang Yan, Kezhi Mao, Fei Shen, and Jinjiang Wang:

‘Machine health monitoring using local feature-based gated recurrent unit networks’, IEEE Transactions on Industrial Electronics, 2017, 65, (2), pp. 1539-1548

[Gu, et al., 2018] Shuqin Gu, Lipeng Zhang, Yuexian Hou, and Yin Song: ‘A position-aware bidirectional attention network for aspect-level sentiment analysis’. In. Proceedings of the 27th International Conference on Computational Linguistics pp. 774-784, 2018

[Goodfellow, et al., 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio: ‘Generative adversarial nets’. In. Advances in neural information processing systems pp. 2672-2680, 2014

[Smolensky, 1986] Paul Smolensky: ‘Information processing in dynamical systems: Foundations of harmony theory’, in Editor (Ed.)^(Eds.): ‘Book Information processing in dynamical systems: Foundations of harmony theory’ (Colorado Univ at Boulder Dept of Computer Science, 1986, edn.), pp.

[Hinton, et al., 2006] Geoffrey E Hinton, and Ruslan R Salakhutdinov: ‘Reducing the dimensionality of data with neural networks’, science, 2006, 313, (5786), pp. 504-507

[Kramer, 1991] Mark A Kramer: ‘Nonlinear principal component analysis using autoassociative neural networks’, AIChE journal, 1991, 37, (2), pp. 233-243

[Heusel, et al., 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter: ‘Gans trained by a two time-scale update rule converge to a local nash equilibrium’. In. Advances in neural information processing systems pp. 6626-6637, 2017

[Torres-Reyes, et al., 2019] Norberto Torres-Reyes, and Shahram Latifi: ‘Audio Enhancement and Synthesis using Generative Adversarial Networks: A Survey’, International Journal of Computer Applications, 2019, 975, pp. 8887

[Zhang, et al., 2016] Yizhe Zhang, Zhe Gan, and Lawrence Carin: ‘Generating text via adversarial training’. In. NIPS workshop on Adversarial Training pp. 2016

[Yu, et al., 2017] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu: ‘Seqgan: Sequence generative adversarial nets with policy gradient’. In. Thirty-First AAAI Conference on Artificial Intelligence pp. 2017

[Wang, et al., 2017] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang: ‘Irgan: A minimax game for unifying generative and discriminative information retrieval models’. In. Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval pp. 515-524, 2017

[Radford, et al., 2015] Alec Radford, Luke Metz, and Soumith Chintala: ‘Unsupervised representation learning with deep convolutional generative adversarial networks’, arXiv preprint arXiv:1511.06434, 2015

[Wang, et al., 2018] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro: ‘Video-to-video synthesis’, arXiv preprint arXiv:1808.06601, 2018

[Wang, et al., 2018] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro: ‘High-resolution image synthesis and semantic manipulation with conditional gans’. In. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 8798-8807, 2018

[Karras, et al., 2019] Tero Karras, Samuli Laine, and Timo Aila: ‘A style-based generator architecture for generative adversarial networks’. In. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 4401-4410, 2019

[Wang, et al., 2019] Huan Wang, Luping Zhou, and Lei Wang: ‘Miss Detection vs. False Alarm:

Adversarial Learning for Small Object Segmentation in Infrared Images'. In. International Conference on Computer Vision (ICCV) pp. 8508-8517, 2019

[Lafferty, et al., 2003] John Lafferty, and Chengxiang Zhai: 'Probabilistic relevance models based on document and query generation': 'Language modeling for information retrieval' (Springer, 2003), pp. 1-10

[Koren, et al., 2009] Yehuda Koren, Robert Bell, and Chris Volinsky: 'Matrix factorization techniques for recommender systems', Computer, 2009, 42, (8), pp. 30-37

[Burges, et al., 2007] Christopher J Burges, Robert Ragno, and Quoc V Le: 'Learning to rank with nonsmooth cost functions'. In. Advances in neural information processing systems pp. 193-200, 2007

[Chen, et al., 2018] Laming Chen, Guoxin Zhang, and Eric Zhou: 'Fast greedy map inference for determinantal point process to improve recommendation diversity'. In. Advances in Neural Information Processing Systems pp. 5622-5633, 2018

[Chae, et al., 2019] Dong-Kyu Chae, Jin-Soo Kang, Sang-Wook Kim, and Jaeho Choi: 'Rating Augmentation with Generative Adversarial Networks towards Accurate Collaborative Filtering'. In. The World Wide Web Conference pp. 2616-2622, 2019

[Wang, et al., 2019] Qinyong Wang, Hongzhi Yin, Hao Wang, Quoc Viet Hung Nguyen, Zi Huang, and Lizhen Cui: 'Enhancing collaborative filtering with generative augmentation'. In. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining pp. 548-556, 2019

[Mirza, et al., 2014] Mehdi Mirza, and Simon Osindero: 'Conditional generative adversarial nets', arXiv preprint arXiv:1411.1784, 2014

[Gao, et al., 2019] Rong Gao, Haifeng Xia, Jing Li, Donghua Liu, Shuai Chen, and Gang Chun: 'DRCGR: Deep Reinforcement Learning Framework Incorporating CNN and GAN-Based for Interactive Recommendation'. In. 2019 IEEE International Conference on Data Mining (ICDM) pp. 1048-1053, 2019

[Chae, et al., 2019] Dong-Kyu Chae, Jung Ah Shin, and Sang-Wook Kim: 'Collaborative adversarial autoencoders: An effective collaborative filtering model under the GAN framework', IEEE Access, 2019, 7, pp. 37650-37663

[Rendle, et al., 2012] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme: 'BPR: Bayesian personalized ranking from implicit feedback', arXiv preprint arXiv:1205.2618, 2012

[Sun, et al., 2020] Changfeng Sun, Han Liu, Meng Liu, Zhaochun Ren, Tian Gan, and Liqiang Nie: 'LARA: Attribute-to-feature Adversarial Learning for New-item Recommendation'. In. Proceedings of the 13th International Conference on Web Search and Data Mining pp. 582-590, 2020

[Kumar, et al., 2019] Sudhir Kumar, and Mithun Das Gupta: '$ c^+ $ GAN: Complementary Fashion Item Recommendation', arXiv preprint arXiv:1906.05596, 2019

[Perera, et al., 2019] Dilruk Perera, and Roger Zimmermann: 'CnGAN: Generative Adversarial Networks for Cross-network user preference generation for non-overlapped users'. In. The World Wide Web Conference pp. 3144-3150, 2019

[Wang, et al., 2019] Cheng Wang, Mathias Niepert, and Hui Li: 'RecSys-DAN: Discriminative Adversarial Networks for Cross-Domain Recommender Systems', IEEE transactions on neural networks and learning systems, 2019, 31, (8), pp. 2731-2740

[Yang, et al., 2014] Haojin Yang, and Christoph Meinel: 'Content based lecture video retrieval using speech and video text information', IEEE Transactions on Learning Technologies, 2014, (2), pp. 142-154

[Lopez, et al., 2017] Glenn Lopez, Daniel T Seaton, Andrew Ang, Dustin Tingley, and Isaac Chuang: 'Google BigQuery for Education: Framework for Parsing and Analyzing edX MOOC Data'. In. Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale pp. 181-184, 2017

[Abdullah, et al., 2016] Amirali Abdullah, Ravi Kumar, Andrew McGregor, Sergei Vassilvitskii, and Suresh Venkatasubramanian: 'Sketching, Embedding and Dimensionality Reduction in Information Theoretic Spaces'. In. Artificial Intelligence and Statistics pp. 948-956, 2016

[Mitra, et al., 2017] Bhaskar Mitra, and Nick Craswell: 'Neural text embeddings for information retrieval'. In. Proceedings of the Tenth ACM International Conference on Web Search and Data Mining pp. 813-814, 2017

[Zhou, et al., 2016] Ningnan Zhou, Wayne Xin Zhao, Xiao Zhang, Ji-Rong Wen, and Shan Wang: 'A general multi-context embedding model for mining human trajectory data', IEEE transactions on knowledge and data engineering, 2016, 28, (8), pp. 1945-1958

[Zhang, et al., 2015] Ziming Zhang, and Venkatesh Saligrama: 'Zero-shot learning via semantic similarity embedding'. In. Proceedings of the IEEE international conference on computer vision pp. 4166-4174, 2015

[Yin, et al., 2017] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze: 'Comparative study of CNN and RNN for natural language processing', arXiv preprint arXiv:1702.01923, 2017

[Zhai, et al., 2018] Zenan Zhai, Dat Quoc Nguyen, and Karin Verspoor: 'Comparing CNN and LSTM character-level embeddings in BiLSTM-CRF models for chemical and disease named entity recognition', EMNLP 2018, 2018, pp. 38

[Papandreou, et al., 2015] George Papandreou, Iasonas Kokkinos, and Pierre-André Savalle: 'Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection'. In. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 390-399, 2015

[Chen, et al., 2017] Zhuyun Chen, and Weihua Li: 'Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network', IEEE Transactions on Instrumentation and Measurement, 2017, 66, (7), pp. 1693-1702

[Devlin, et al., 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova: 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) pp. 4171-4186, 2019

[Peters, et al., 2018] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer: 'Deep Contextualized Word Representations'. In. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) pp. 2227-2237, 2018

[Mikolov, et al., 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean: 'Efficient estimation of word representations in vector space', arXiv preprint arXiv:1301.3781, 2013

[Roark, et al., 2007] Brian Roark, Murat Saraclar, and Michael Collins: 'Discriminative n-gram language modeling', Computer Speech & Language, 2007, 21, (2), pp. 373-392

[Valverde-Albacete, et al., 2014] Francisco J Valverde-Albacete, and Carmen Peláez-Moreno: '100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox', PloS one, 2014, 9, (1), pp. e84217

[Pennington, et al., 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning: 'Glove:

Global vectors for word representation'. In. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) pp. 1532-1543, 2014

[Deerwester, et al., 1990] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman: 'Indexing by latent semantic analysis', Journal of the American society for information science, 1990, 41, (6), pp. 391-407

[Devlin, et al., 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova: 'Bert: Pre-training of deep bidirectional transformers for language understanding', arXiv preprint arXiv:1810.04805, 2018

[Paszke, et al., 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga: 'Pytorch: An imperative style, high-performance deep learning library'. In. Advances in neural information processing systems pp. 8026-8037, 2019

[Gers, et al., 1999] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins: 'Learning to forget: Continual prediction with LSTM', 1999

[Cho, et al., 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio: 'Learning phrase representations using RNN encoder-decoder for statistical machine translation', arXiv preprint arXiv:1406.1078, 2014

[Neubig, et al., 2013] Graham Neubig, and Kevin Duh: 'How Much Is Said in a Tweet? A Multilingual, Information-theoretic Perspective'. In. AAAI Spring Symposium: Analyzing Microtext pp. 32-39, 2013

[Pazzani, 1999] Michael J Pazzani: 'A framework for collaborative, content-based and demographic filtering', Artificial intelligence review, 1999, 13, (5-6), pp. 393-408

[Lin, et al., 2020] Jiayin Lin, Geng Sun, Jun Shen, David Pritchard, Tingru Cui, Dongming Xu, Li Li, Ghassan Beydoun, and Shiping Chen: 'Deep-Cross-Attention Recommendation Model for Knowledge Sharing Micro Learning Service'. In. International Conference on Artificial Intelligence in Education pp. 168-173, 2020

[Dobozy, 2017] Eva Dobozy: 'University lecturer views on pedagogic lurking'. In. 2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT) pp. 1-2, 2017

[Beaudoin, 2002] Michael F Beaudoin: 'Learning or lurking?: Tracking the "invisible" online student', The internet and higher education, 2002, 5, (2), pp. 147-155

[Dennen, 2008] Vanessa Paz Dennen: 'Pedagogical lurking: Student engagement in non-posting discussion behavior', Computers in Human Behavior, 2008, 24, (4), pp. 1624-1633

[Paszke, et al., 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga: 'PyTorch: An imperative style, high-performance deep learning library'. In. Advances in Neural Information Processing Systems pp. 8024-8035, 2019

[Zhou, et al., 2019] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai: 'Deep interest evolution network for click-through rate prediction'. In. Proceedings of the AAAI Conference on Artificial Intelligence pp. 5941-5948, 2019

[Yu, et al., 2020] Jifan Yu, Gan Luo, Tong Xiao, Qingyang Zhong, Yuquan Wang, Junyi Luo, Chenyu Wang, Lei Hou, Juanzi Li, and Zhiyuan Liu: 'MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs'. In. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics pp. 3135-3142, 2020

[Tallapally, et al., 2018] Dharahas Tallapally, Rama Syamala Sreepada, Bidyut Kr Patra, and Korra Sathya Babu: 'User preference learning in multi-criteria recommendations using stacked auto encoders'. In. Proceedings of the 12th ACM Conference on Recommender Systems pp. 475-479, 2018

[Lin, et al., 2020] Jiayin Lin, Geng Sun, Jun Shen, Tingru Cui, David Pritchard, Dongming Xu, Li Li, Wei Wei, Ghassan Beydoun, and Shiping Chen: 'Attention-Based High-Order Feature Interactions to Enhance the Recommender System for Web-Based Knowledge-Sharing Service'. In. International Conference on Web Information Systems Engineering pp. 461-473, 2020

[Feng, et al., 2019] Yufei Feng, Fuyu Lv, Weichen Shen, Menghan Wang, Fei Sun, Yu Zhu, and Keping Yang: 'Deep session interest network for click-through rate prediction', arXiv preprint arXiv:1905.06482, 2019

[Wang, et al., 2017] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang: 'Deep & cross network for ad click predictions'. In. Proceedings of the ADKDD'17 pp. 1-7, 2017

[Schröder, et al., 2011] Gunnar Schröder, Maik Thiele, and Wolfgang Lehner: 'Setting goals and choosing metrics for recommender system evaluations'. In. UCERSTI2 workshop at the 5th ACM conference on recommender systems, Chicago, USA pp. 53, 2011

[Schafer, et al., 2007] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen: 'Collaborative filtering recommender systems': 'The adaptive web' (Springer, 2007), pp. 291-324

[Li, et al., 2017] Xiaopeng Li, and James She: 'Collaborative variational autoencoder for recommender systems'. In. Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining pp. 305-314, 2017

[Liu, et al., 2018] Yu Liu, Shuai Wang, M Shahrukh Khan, and Jieyu He: 'A novel deep hybrid recommender system based on auto-encoder with neural collaborative filtering', Big Data Mining and Analytics, 2018, 1, (3), pp. 211-221

[Lee, et al., 2017] Wonsung Lee, Kyungwoo Song, and Il-Chul Moon: 'Augmented variational autoencoders for collaborative filtering with auxiliary information'. In. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management pp. 1139-1148, 2017

[Koren, 2008] Yehuda Koren: 'Factorization meets the neighborhood: a multifaceted collaborative filtering model'. In. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining pp. 426-434, 2008

[Glorot, et al., 2010] Xavier Glorot, and Yoshua Bengio: 'Understanding the difficulty of training deep feedforward neural networks'. In. Proceedings of the thirteenth international conference on artificial intelligence and statistics pp. 249-256, 2010

[Houle, et al., 2010] Michael E Houle, Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek: 'Can shared-neighbor distances defeat the curse of dimensionality?'. In. International conference on scientific and statistical database management pp. 482-500, 2010