Title	Content Generation and Serious Game Implementation for Security Awareness Training
Author(s)	曽，由美子
Citation	
Issue Date	2021-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/17105
Rights	
Description	Supervisor: Razvan Beuran，先端科学技術研究科，修士（情報科学）

Master's Thesis

Content Generation and Serious Game Implementation for Security Awareness Training

1910129 ZENG Youmeizi

Supervisor	Research Associate Professor Razvan Beuran
Main Examiner	Research Associate Professor Razvan Beuran
Examiners	Professor Yasuo Tan
	Associate Professor Yuto Lim
	Research Associate Professor Ken-ichi Chinen

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

March 2021

Abstract

With the growth of global informatization, the extensive application of information technology and the widespread use of intelligent terminals, the Internet has penetrated every aspect of our lives, and has increasingly become an indispensable part of our daily existence. However, while we use the Internet to communicate, do online shopping and so on, hence it brings infinite convenience to people, we cannot ignore the associated cybersecurity risks.

In 2020, the global outbreak of COVID-19 began. To prevent the spread of the virus, people began to reduce social activities and maintain social distancing. Many governments and companies began to implement remote work measures. However, the remote work increased the cybersecurity risks to organizations. Cybercriminals use phishing emails related to COVID-19 to flood employees' inboxes, and seemingly harmless attachments are malicious software that lures unsuspecting employees to open them.

Such cyberattacks bring economic losses to companies and organizations, and can be used to gather information for political motives, or to cause people panic or fear. However, cybersecurity incidents are not only caused by system vulnerabilities. According to a survey by IBM, human factors are the weakest link in cyber defense strategies, and about 95% of cybersecurity risks are due to human errors.

No one can avoid all the mistakes, but companies or organizations can try to effectively avoid security incidents caused by human error, and reduce the potential risks and losses by training employees on cybersecurity awareness. Individuals also need to increase their security awareness in order to prevent various cyberattacks, and to ensure that their rights are not violated.

There are many methods to conduct training on cybersecurity awareness. In traditional ways, we will learn in the classroom or through reading materials. However, those traditional learning strategies often give learners a "dry" and "boring" learning experience, which will lead them to reduce their motivation to learn more about subject contents. Although learning by watching videos can reduce the "dry" part, it still lacks interactivity and practicality.

Compared to the above training methods, this research proposes to use serious games to conduct training on cybersecurity awareness. Serious games have many potential advantages, such as flexibility, interactivity, low cost-effectiveness, and low risk. Besides, the most attractive advantage is that learners can repeatedly play the same serious game to explore the different results caused by different actions, even if such results may have

a disastrous impact in real life.

Each pedagogic training method brings different expected effects, but these effects also depend on the actual education or training content. Creating this content is indeed one of the most time-consuming and labor-intensive tasks that developers face when designing a teaching and training program.

Developers will typically ask professionals in related fields to design customized content so as to ensure the quality of the instructional content. As the risks related to Internet increase, there will be new related knowledge that needs to be understood at any time. The previous method to generate content cannot satisfy learners' expectations for a large amount of new education content. Therefore this research proposes to use Natural Language Generation (NLG) to automatically generate the training content. In particular, we used Naive Bayes models to generate cybersecurity training content for the platform presented in this thesis.

Before generating the content, we need to prepare the dataset. As training data, we extracted the paragraphs, sentences (containing the answer), questions and answers in SQuAD1.1. Then we preprocessed and standardized the data to eliminate human error or incorrectness, and avoid the impact of repeated data on the results. As actual prediction data for the platform, we extracted 2640 cybersecurity concepts from DBpedia by using "computer security" as keyword, and collected 2315 concept definitions from Wikipedia for the above concepts. Since the original data cannot be used directly, we performed feature engineering to select the key features in the text and encode them, and to convert them into data that can be used for machine learning. After feature engineering, some methods were used to deal with imbalanced data, thus prevent the dominance of larger data sets. In the end we divided the final processed data into 80% as training data and 20% as test data.

The training data was used to train Naive Bayes models, and the test data to provide an unbiased evaluation of the trained model. By using 9 evaluation metrics and tuning the parameters, we finally selected the SMOTE method to train the Bernoulli Naive Bayes model after performing isotonic calibration. The prepared prediction data was inserted into this trained model, then used to generate cloze question and answer pairs. We combine and stored all the prediction data and collected data in the form of a database of training content.

After solving the problem of creating training content, we developed a web application, named CyATP (Cybersecurity Awareness Training Platform), to display the generated content as a convenient way to conduct security awareness training. This application's front end mainly uses the open source framework Bootstrap, and jQuery was used to design the web pages. The back end uses the lightweight python web framework Flask. The dataset of keywords and concept maps are stored in the relational database Neo4j, and the generated questions and puzzle data are stored in a JSON file.

The CyATP platform is roughly divided into two parts: the learning activity component and the serious game component. The learning activity component includes two pages: Concept Map and Learn Concepts. Trainees use the web interface to access those two pages, and learn about the security concepts they want to understand through exploratory interest learning. The serious game component also includes two

pages: Take Quiz and Crossword Puzzle. Trainees play those games to test and deepen their knowledge.

We recruited some volunteers to use our platform for training and asked them to fill out questionnaires after using it. The trainees gave feedback according to their level of agreement with the statements we provided about CyATP. Each question was graded from 1 (strongly disagree) to 5 (strongly agree). The questionnaire was used to evaluate the quality of the generated content, the usability of the platform, and the serious game component of the platform.

The trainees' evaluation of the concept map based learning content produced a very high score, and the general opinion was that the concept text is easy to understand and suitable for learning. We also used the SUS (System Usability Scale) to evaluate the usability of the CyATP platform. According to the average score of 80.5 given by the trainees, CyATP is a good and acceptable platform for cybersecurity awareness training. For the evaluation of serious games, we use 9 factors and 29 items. The trainees' evaluation shows that those serious games are easy to use, give users immediate feedback, have clear goals, and it is efficient to learn security knowledge while playing the games.

The implementation of the CyATP cybersecurity awareness training platform is a significant contribution of this research. CyATP is a tool for everyone who wants to gain or expand their knowledge in cybersecurity awareness. By exploratory interest learning and serious games to enhance their interest, learners can increase their security awareness knowledge and put it to use in their daily life. CyATP also provides a versatile platform for security educators, who can generate additional customized training content, then use the already-built web application structure to conduct training activities.

Keywords: Security awareness training, Content generation, Serious game

Contents

List of Figures

List of Tables

Abbreviation

ADASYN	Adaptive Synthetic
AI	Artificial Intelligence
API	Application Programming Interface
AR	Augmented Reality
AUC	Area Under the Curve
CDN	Content Delivery Network
CTF	Capture the Flag
CyATP	Cybersecurity Awareness Training Platform
DEP	Syntactic Dependency
DL	Deep Learning
DNS	Domain Name System
FMA	Free Music Archive
GIF	Graphics Interchange Format
HTML	Hypertext Markup Language
HTTPS	Hypertext Transfer Protocol Secure
JPEG	Joint Photographic Experts Group
JSON	Javascript Object Notation
LOD	Linked Open Data
MAP	Maximum A Posteriori Estimation
ML	Machine Learning
MNIST	Modified National Institute of Standards and Technology database

NB	Naive Bayes
NER	Named Entity Recognition
NIH	National Institutes of Health
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
POS	Part-of-Speech Tag
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Oversampling Technique
SQuAD	Stanford Question Answering Dataset
SSL	Secure Sockets Layer
SUS	System Usability Scale
SVM	Support Vector Machine
TAG	Detailed Part-of-Speech Tag
TCP	Transmission Control Protocol
TTFB	Time To First Byte
WHO	World Trade Organization
WSGI	Web Server Gateway Interface

Acknowledgment

Time flies, two years of graduate school life is coming to an end, and I can still remember when I stepped into the JAIST campus. No matter where I am, I will not forget this wonderful time here in the following days.

Foremost, I would like to express my sincere gratitude to my supervisor Razvan Beuran, who helped me with many things in the research. From idea to plan, then implementation to realization, he patiently guided me and gave me suggestions at every stage, witnessing my growth in research. Without his help, my research and writing of this thesis would not be completed so smoothly. I could not have imagined having the best mentor in the study.

Second, I want to thank my friends for their continued support and encouragement. Especially those who have been by my side and fighting with me. It was you who made my college life rich and colorful in the past two years and left very precious memories.

Last but not least, I want to thank my family who have always supported me. Thank you for believing and supporting my choice and let me do what I want to do. Even if you don't understand the research I've done, you are still doing your best to help and solve my difficulties. There are always difficult things in life that make me feel sad and depressed, but you always tell me that there is nothing impossible.

Thanks to everyone I met, I became who I want to be because of you. Treasure all the memories, and see you again.

Chapter 1

Introduction

In this chapter, we first talk about the importance of cybersecurity awareness training and the problems encountered in current training, propose three issues that cannot be ignored. Next, we introduce the contributions of this thesis. Finally, we describe the structure of the thesis.

1.1 Motivation

With the growth of global informatization, extensive application of information technology and widespread use of intelligent terminals, the Internet has penetrated every aspect of our lives and has increasingly become an indispensable part of our daily lives. However, while we use the Internet to communicate, shopping, and enjoy it brings people infinite convenience, we cannot ignore the cyberattacks and risks.

When the global outbreak of the COVID-19 in 2020, it threatens everyone, organizations and government agencies worldwide and even came to a standstill. In order to avoid the spread of the virus, many governments and companies have recommended measures to the remote workforce, which is followed by many cyber challenges that companies cannot afford to ignore. Although the virus has a great impact on our lives, cybercrime has not slackened and has targeted unsuspecting individuals and organizations to steal personal information or company data.

Based on the study by Ponemon Institute [1], 2,215 IT and security workers in the US, the UK and other countries participate in the survey on how organizations' cybersecurity has been affected by the move to telework. Since the virus outbreak, 63% of U.S. respondents have seen an increase in phishing/social engineering attacks, but unfortunately, 50% of U.S. respondents said their organization does not provide cybersecurity training for remote workers.

Before the coronavirus broke out, cybersecurity personnel tried to keep up with the pace of cyberattacks and provide defenses, but COVID-19 increased their burden. Most of the cyberattacks during this period were phishing, conference bombing, and ransomware. To defend against these attacks, in addition to maintaining vulnerabilities, it is more important to train individuals on cybersecurity awareness.

In the current social situation, traditional training methods in the classroom or reading training materials have not met the demand. We want to find a combined education and entertaining way to conduct cybersecurity awareness training and improve learning motivation. Simultaneously, like training, the impact of learning content on trainers is not ignored, and many quality materials are needed.

For security awareness training, the following three issues cannot be ignored: (i) Firstly, how to quickly and efficiently obtain a large amount of customizable training content; (ii) Secondly, how to base on the training materials, provide a cybersecurity awareness training platform for learners to combine education and entertaining; (iii) Finally, how to enable security trainers to easily build their own platform according to their needs and implement security awareness training. This research aims to solve these three questions.

1.2 Contributions

The following points can be considered as main contributions of this research:

- We propose a way to automatically generate cybersecurity training content using Natural Language Generation technology. This method can quickly, easily and efficiently generate a large amount of training content, can meet users' needs.

- The proposed method was highly evaluated. Thus, for user evaluation with the generated content, the average score of the result was 4.07. And the evaluation results of the trained model used to generate the content had a high accuracy score of 84.3%.

- We develop and implement a cybersecurity awareness training platform CyATP. It provides a tool for everyone who wants to gain or expand their knowledge in cybersecurity awareness.

- The implemented platform was also highly evaluated. Thus, the performance evaluation showed that it is fast, and the browsing experience is smooth. The user evaluation via the System Usability Score (SUS) resulted in an average score of 80.5, which means a good and acceptable platform.

- We provide the source code of CyATP as an open-source project on GitHub that can be easily deployed by security educators and used for training content generation and as a training platform.

1.3 Structure of Thesis

The remainder sections of this thesis are organized as follows:

- Chapter 2 – Research Background: We introduce the current cybersecurity training method and proposed to use serious games with potential advantages for training.

Then we discuss the content generation problem that cannot be ignored in cybersecurity training, and proposed to use NLG technology to automatically generate security training content.

- Chapter 3 – Training Content Generation: We describe the details of training content generation in this research. First, we introduce an overview of content generation, then we discuss data preparation and training models, and finally we use the trained models to make predictions to generate the final training content.
- Chapter 4 – Cybersecurity Awareness Training Platform: In order to display and utilize the generated content, we have built a web application platform CyATP for trainees. In this section we talk about the framework and implementation of this platform and introduce the specific functions of each page.
- Chapter 5 – Evaluation: We invited some volunteers to use our platform, and then evaluated our research from three aspects: the quality of generated training content, the usability of the platform, and the serious games, which are presented in this section.
- Chapter 6 – Conclusion and Future Work: We summarize this thesis's whole work and give suggestions about aspects that could be improved in the future.
- Appendix: The appendix provides information related to the questionnaire survey, including an introduction to the questionnaire, the implementation process, and the specific questions used in the survey.

Chapter 2

Research Background

In this chapter, we focus on introducing the background knowledge needed in this thesis. First, a brief introduction to cybersecurity awareness training situation and method. Then, we talk about content generation and using the Natural Language Generation (NLG) method for content generation. Finally, introduce the definition and application of serious games, and compare serious games used in cybersecurity awareness training.

2.1 Cybersecurity Awareness Training

As information technology develops, and the widespread application of information technology, the Internet has increasingly become an indispensable part of our lives. However, while information technology brings unlimited convenience and benefits to people, it also brings risks. For example: On June 8, 2020, Japanese automobile production company Honda was attacked by the ransomware "EKANS." The attack caused its factories' production and shipment system in Japan and overseas to suspend operation, bringing substantial economic losses to the company. Of course, security incidents do not only occur in one region but on a global scale. In the global medical organizations fight the COVID-19, about 25000 email addresses and passwords were leaked online, which belonged to the World Trade Organization (WHO), the National Institutes of Health (NIH), and the Gates Foundation.

Cyberattacks will bring economic losses to the company and organization, and can be used gather information for political motives, or to cause people panic or fear. However, cybersecurity incidents are not only caused by system vulnerabilities. According to a survey by IBM [2], human factors are the weakest link in cyber defense strategies, and about 95% of cybersecurity risks are due to human errors.

There are broadly two types of human errors: skills-based errors and decision-based errors. The main difference between the two human errors is whether the operator has the required knowledge, and does the correct action. Skill-based errors such as mistakes, the operator know the correct operation method, but the mistakes lead to errors. The reason for errors happens maybe the operator is tired or distracted. Decision-based errors is the wrong decision made by the operator. It usually includes operators who lack the necessary knowledge or unawareness.

No one can avoid all the mistakes, but companied or organizations can try to effectively avoid security incidents caused by human error, and reduce the potential risks and losses by training employees on cybersecurity awareness. Individuals also need to increase their security awareness in order to prevent various cyberattacks, and to ensure that their rights are not violated.

2.1.1 Cybersecurity Awareness Training Methods

While we use information technology, there are also many potential dangers. Many companies organize their employees to learn cybersecurity knowledge to prevent data leakage caused by human error; many school organizations train students' cybersecurity awareness to prevent them from becoming the victim of some cyberattack.

There are many methods to conduct training on cybersecurity awareness. In traditional ways, we will learn in the classroom or through reading materials. However, those traditional learning strategies often give learners a "dry" and "boring" learning experience, which will lead them to reduce their motivation to learn more about subject contents. Although learning by watching videos can reduce the "dry" part, it still lacks interactivity and practicality.

Using hands-on training (such as the apprentice model), not only allows learners to apply their knowledge to real-world situations, but also provides the learners with reinforcement and feedback. Nevertheless, this training method is used in the real world, which may bring irreparable risk.

CTF (Capture the Flag) is widely used for cybersecurity competitions and awareness training. Player teams can solve various security problems of different complexity in a limited time ranging from hours to days. The participates usually asked to solve the tasks and find a specific piece of text that may be hidden in files or images [3]. In these challenges, the only clear goal is to find the flag (like a string), which may give clues in some competitions.

However, the CTF might discourage some learners, especially beginners. The tasks are usually too difficult for less experienced participants. Sometimes, without guidance, novice learners miss essential learning goals and take longer to learn concepts. Moreover, the level of tasks must be designed by professionals, which will cost a lot of money. Some competitions only pay attention to whether the final result finds the flag, not the finding process.

Compared to the above training methods, this research proposes to use serious games to conduct training on cybersecurity awareness. Serious games have many potential advantages, such as flexibility, low cost-effectiveness, low risk, and standardized assessments that can be compared between learners. Besides, the more attractive advantage is that learners can repeatedly play the same serious game to explore the different results of different actions, even if such results may have a disastrous impact on real life.

The table 2.1 on the following page shows a comparison between classroom learning, reading material learning, watching video learning, hands-on training, CTF, and serious game training.

Comparison Item [4]	Classroom learning	Reading material learning	Watching video learning	Hands-on training	CTF (Capture the Flag)	Serious game training
Flexibility	✗	✓	✓	✓	✓	✓
Interactive	✓	✗	✗	✓	✓	✓
Immediate feedback	✗	✗	✗	✓	✓	✓
Personalized learning	✗	✗	✗	✓	✓	✓
Increases motivation	✗	✗	✗	✓	✓	✓
Application to real world environment	✗	✗	✗	✓	✓	✓
Low physical risk	✓	✓	✓	✗	✓	✓
Cost effective	✗	✓	✓	✗	✗	✓
Standardized assessment	✓	✗	✗	✗	✗	✓

Table 2.1: Comparison of 6 learning and training ways in cybersecurity awareness training.

2.2 Content Generation

Each cybersecurity pedagogic training method brings different expected effects, but these effects also depend on the actual education or training content. Careful production of content is necessary because the content has a considerable influence on learners. Creating this content is indeed one of the most time-consuming and labor-intensive tasks that developers face when designing a teaching and training program.

Usually, developers will ask professionals in related fields to design customized content to ensure the quality of the instructional content. As the risks related to Internet increase, there will be new related knowledge that needs to be understood at any time. The previous method to generate content cannot satisfy learners' expectations for a large amount of new education content.

How to efficiently and quickly generate a large amount of educational content has become a key point of cybersecurity awareness training and education. Not only the content itself, but the way it is produced also affects the quality and cost of education. If the generated content is not timely enough or takes a long period, when it is found that the content is not suitable for the existing teaching methods, it may be too late to regenerate. Especially if this problem is discovered in the later development stages, it may cause expensive additional work. Therefore, the method of content generation is an important topic.

In recent years, with the development of Artificial Intelligence (AI) technology and the increase in types, Natural Language Processing (NLP) technology has become widely available. It has evolved from a system based on simple templates and rules to can understand complex human grammar. In the past, we may be dissatisfied with the ridiculous content and inconsistent results generated by machine translation and Natural Language Processing. However, through researchers' continuous efforts in this field, AI has become more reliable and mature. In the figure 2.1, we can know NLP is a subset of AI, and uses ML (Machine Learning) and DL (Deep Learning) technologies.

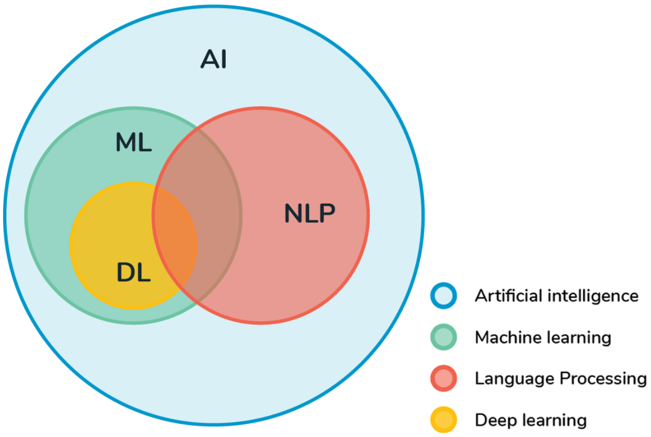


Figure 2.1: The relationship between AI, ML, DL and NLP [5].

There have been many examples of using NLP to generate content. For example, Heliograf, an artificial intelligence robot independently developed by The Washington Post, issued about 300 reports during the Rio Olympics. Chatbots are now also used by many companies, that can provide customers with consultation, complaints and help with related procedures. The development of these chatbots not only requires advanced NLP capabilities to understand customers' needs, but also requires NLG capabilities to answer customers' questions.

2.2.1 Natural Language Generation (NLG)

Natural Language Generation (NLG) is focused on producing human understandable natural language output from some nonlinguistic information as inputs. Generally, the goal of the NLU systems uses knowledge about language and the application domains to automatically generate documents, reports, instructions, help messages, and other types of texts [6].

From the figure 2.2 we can know the NLG and NLU is are the subset of NLP.

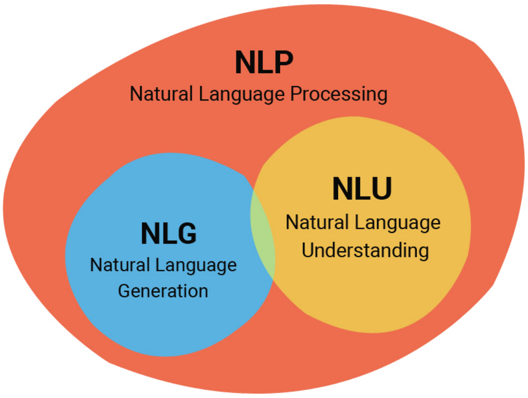

NLP
Natural Language Processing
NLU
Natural Language Understanding
NLG
Natural Language Generation

Figure 2.2: The relationship between NLP, NLG and NLU [7].

NLG and NLU are very closely. They both study computer systems that understand human language, share many of the same theoretical foundations, and are often used together in application programs [6]. For example, conversational AI robots like Siri use NLU and NLG technologies to communicate with people. However, considering the process, NLG is the inverse of NLU. NLG is the process of mapping from computer structured data to human language, whereas NLU is the process of mapping human language to computer structured data. Compared with NLG, NLU is more difficult to implement. Because the language itself is ambiguous and complex, sometimes it is more necessary to understand it better through the context. This feature brings challenges to construct the NLU system.

Question Generation from text is an NLG task concerned with generating questions from unstructured text [8]. In this research, we use NLG to generate the cloze question to provide the cybersecurity learning content.

2.2.2 Content Generation Using NLG

NLG can generate content, but to automate this process and extract accurate data, Machine Learning is required. Machine Learning uses computer algorithms to analyze data and make intelligent decisions based on what has been learned without being explicitly programmed. The algorithm teaches the machine how to automatically learn and improve from experience, accelerate basic text analysis functions, and ultimately convert unstructured text into usable data.

Machine Learning can be divided into three types: supervised learning, unsupervised learning, and reinforcement learning. Supervised Machine Learning uses labeled or tagged datasets to train algorithms, and the trained model uses learned experience to classify data or accurately predict the output. Also, there are many algorithms for supervised NLP Machine Learning, such as the Naive Bayes Model, Decision Tree, K-Nearest Neighbor, Random Forest, Logistic Regression, and Support Vector Machine (SVM). Unsupervised learning is to classify the original material in order to understand its internal structure. At the beginning of learning, it did not know the classification results were correct, and it took the initiative to find out the rules of its potential categories from these materials. The algorithms like Clustering, Dimensionality Reduction, and so on. Reinforcement learning is that the agent learns in a "trial and error" manner, to get the maximize reward by interacting with the environment. Compared with supervised learning, reinforcement learning is no need for the labeled pairs and explicitly correct sub-optimal actions.

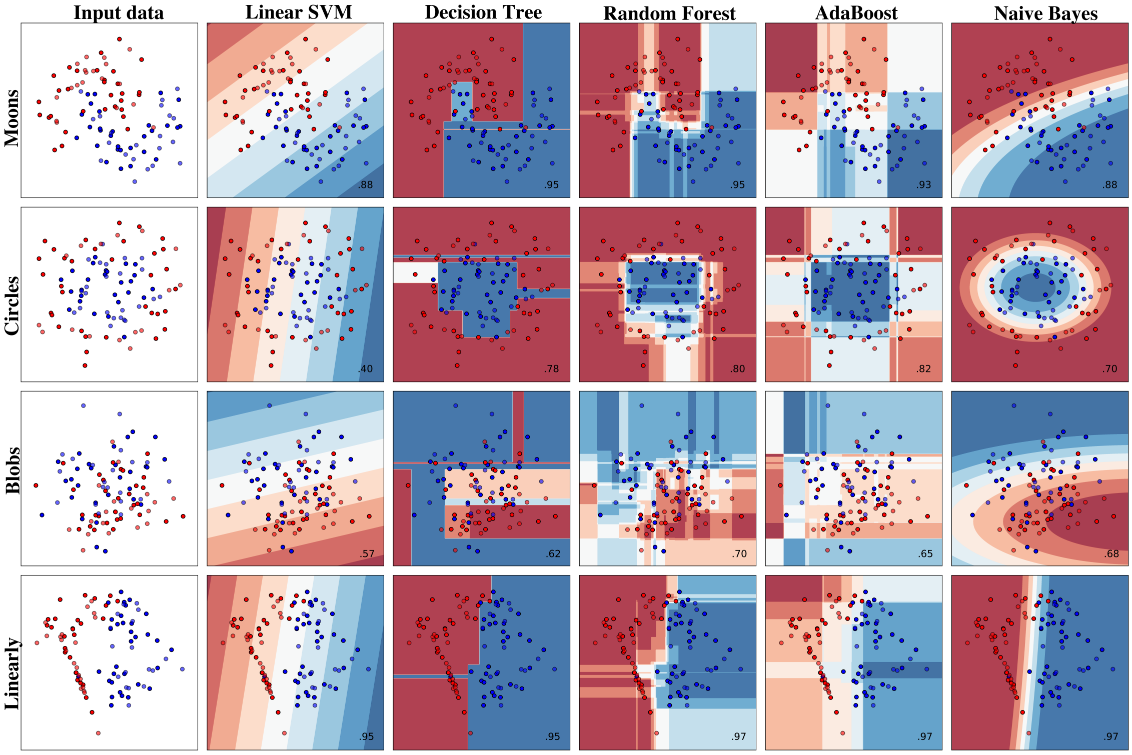


Figure 2.3: Compare the performance of 5 models on 4 types of data.

We used four synthetic data sets (moon, circles, blob, and linearly) in scikit-learn [9] to test the performance of Linear Support Vector Machine (SVM), Decision Tree, Random Forest, AdaBoost, and Naive Bayes. Form the results of the figure 2.3 on the previous page, on the moons data set, the accuracy of Decision Tree and Random Forest is relatively high; on the circle data set, Adaboost performs best; on the blobs data set, the Random Forest can reach up to 70% accuracy; on the linear data set, Random Forest, AdaBoost and Naive Bayes have the highest accuracy.

In this research, cybersecurity awareness training materials include keywords, concept maps, quizzes, and crossword puzzles. In the quiz part, the questions and answers require Natural Language Generation. Generate the cloze question is a linear classification problem. According to the results in figure 2.3 on the preceding page, Random Forest, AdaBoost and Naive Bayes have the highest accuracy on the linear data sets. Among those three machine learning method, Naive Bayes model is good at dealing with classification problems, such as the classification of spam, also the training time is short, and the performance can be stable even using few data for training. Therefore we proposes to use Naive Bayes models to generate cybersecurity training content.

Bayes' Theorem

Bayes' theorem gives a method to calculate the posterior probability P(A|B), through P(A), P(B) and P(B|A). The equation shows below:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

The Naive Bayes model is based on Bayes' theorem. Given the value of a class variable, the "naive" assumption of conditional independence between each pair of features is applied. According to the derivation, the formula is finally expressed as:

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i \mid y)$; the former is then the relative frequency of class (y) in the training set.

The main difference between different naive Bayes methods lies in their assumptions about the distribution of $P(x_i \mid y)$.

Although the independence assumption of Naive Bayes method is usually incorrect in real life, it works well in many practical situations (suck as document classification and spam filtering).

Types of Naive Bayes Algorithms

According to the predicted data distribution, the Naive Bayes method is divided into Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes and Complement Naive Bayes.

The Gaussian Naive Bayes model assumes that the features may be normally distributed, the conditional probability $P(x_i \mid y)$ will change in the following manner (The parameters σ_y and μ_y are estimated using maximum likelihood.):

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

The Multinomial Naive Bayes assumes that the data is a multivariate distribution, and this classic model is often used in text classification. The conditional probability shows below:

$$P(x_i \mid y) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i p_{k_i}^{x_i}$$

The Bernoulli Naive Bayes assumes that the data is a multivariate Bernoulli distribution, each feature is a binary variable, and allowing multiple features. The formulate shows below:

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)$$

The Complement Naive Bayes was designed to correct the "severe assumptions" made by the Multinomial Naive Bayes, and is an adaptation of the Multinomial Naive Bayes. It uses statistics from the complement of each class to compute the model's weights. The procedure for calculating the weights show below:

$$\hat{\theta}_{ci} = \frac{\alpha_i + \sum_{j:y_j \neq c} d_{ij}}{\alpha + \sum_{j:y_j \neq c} \sum_k d_{kj}}$$

$$w_{ci} = \log \hat{\theta}_{ci}$$

$$w_{ci} = \frac{w_{ci}}{\sum_j |w_{cj}|}$$

We also use the four synthetic data sets (moon, circles, blob, and linearly) in scikit-learn to test the performance of Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes and Complement Naive Bayes. The four Naive Bayes models' performance results are shown in figure 2.4 on the next page.

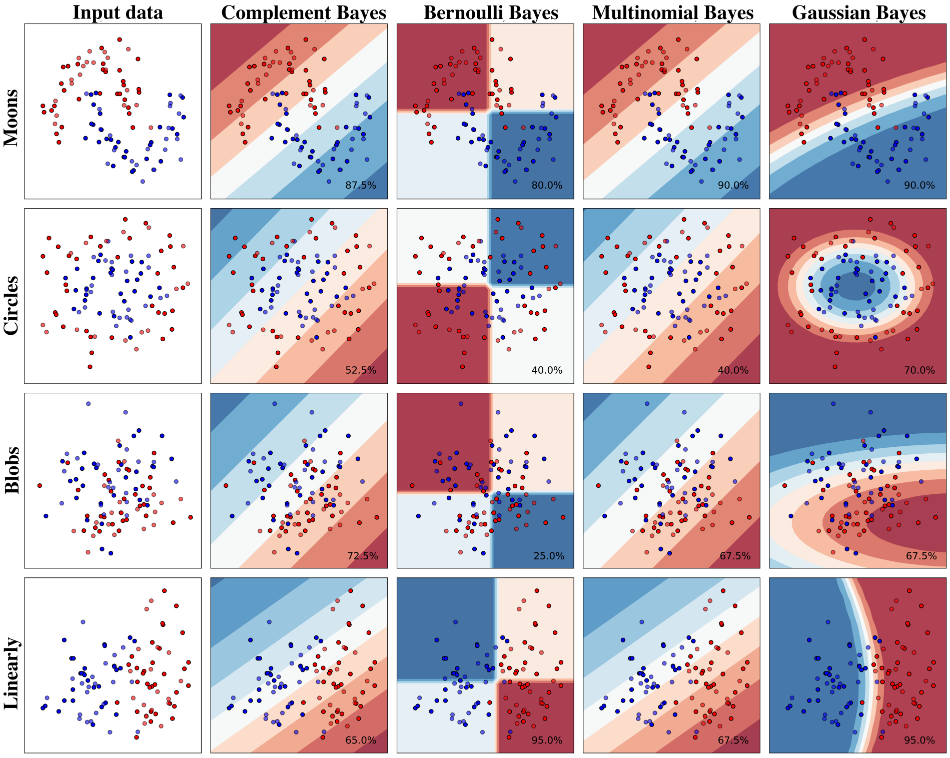


Figure 2.4: Compare the performance of 4 Naive Bayes models on 4 types of data.

Form the results of the figure 2.4, on the moons data set, the accuracy of Gaussian Naive Bayes and Multinomial Naive Bayes is relatively high; on the circle data set, Gaussian Naive Bayes performs best; on the blobs data set, the Complement Naive Bayes can reach up to 72.5% accuracy; on the linear data set, Gaussian Naive Bayes and Beinoulli Naive Bayes have the highest accuracy 95%.

However, the synthesized data set cannot represent the real data set, we will select which model is suitable for cybersecurity awareness training content generation in Section 3.3.

2.3 Serious Games

The games have been attracting people since they came out, and most people spent a lot of leisure time on them. But the games also a tool for the education field, such as the serious games. In this section, will introduce the definition of the serious games, the features of the serious games, and the serious games in cybersecurity awareness training.

2.3.1 Definition

Abt [10] first introduced the "Serious Game" in its modern meaning in 1970. One of the goals of his research is to use serious games for education and training. The serious games T.E.M.P.E.R. designed by Abt used to study the Cold War conflict for military officers on a world-wide scale. He gives a clear definition of "Serious Game":

> "Games may be played seriously or casually. We are concerned with serious games in the sense that these games have an explicit and carefully thought-out educational purpose and are not intended to be played primarily for amusement. This does not mean that serious games are not, or should not be, entertaining."

Also, many researchers and professionals have redefined the "Serious Game." Sawyer created the Serious Game Initiative in 2002, and his white paper [11] suggests using the technology and knowledge from the entertainment video game industry to improve game-based simulations in public policy.

However, the definition of the development is still has debated. A popular definition by Zyda [12] in 2005:

> "A mental contest, played with a computer in accordance with specific rules, that uses entertainment, to further government or corporate training, education, health, public policy, and strategic communication objectives."

It is also worth mentioning the fact that while serious games did not originally refer to video games alone, those extend to any other type of games.

In addition, it is also necessary to clarify the difference between serious games and gamification. From the Deterding's paper [13], we can see the figure 2.5 on the next page, the serious game is more inclined to the two elements of the "whole" and "game", gamification is more inclined to the two elements of the "elements" and "game."

Deterding also defined the "serious games" as

> Using the designed game as the whole game with goals related to non-entertainment purposes, and the design for playful interactions.

In this research, we more simply defined "serious game" as

> "Games which incorporate pedagogic elements."

Furthermore, its purpose is to emphasize the importance of the whole game, which is a feature that cannot be ignored in serious games.

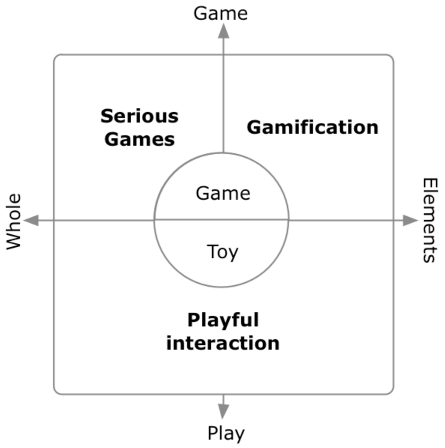
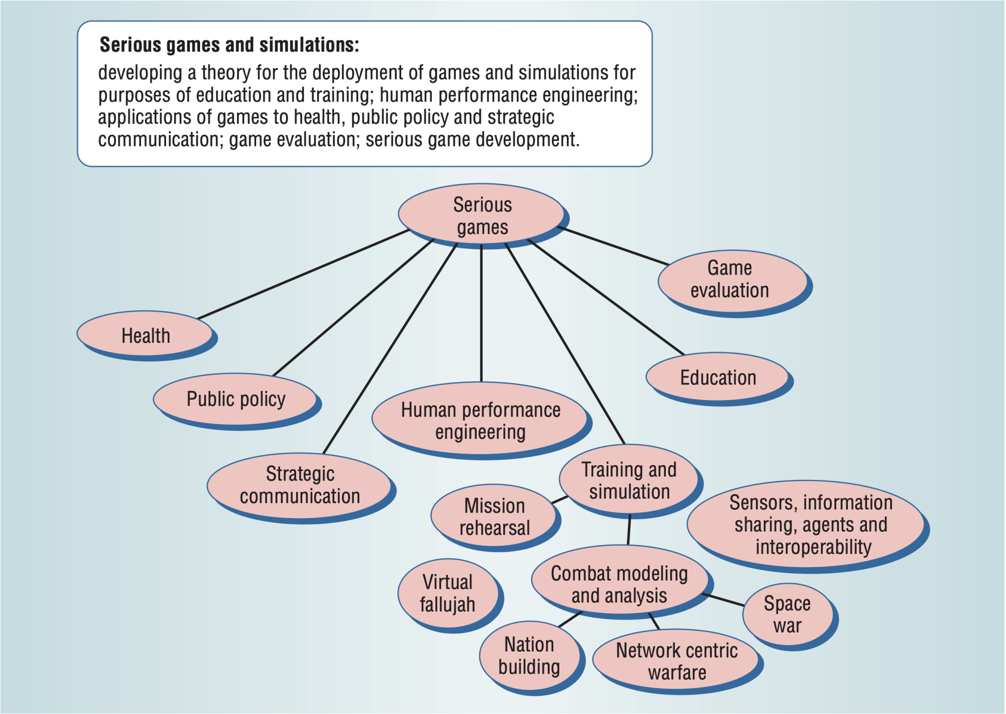


Figure 2.5: Difference between gamification, serious games and playful interaction [13].

2.3.2 Features

Serious games are now developing rapidly and have become a hot topic of research. They are used in many different fields since they can be applied to a broad range of problems and challenges. As figure 2.6 shows, serious games be used in healthcare, public policy, training and education, game evaluation, and so on.



Serious games and simulations: developing a theory for the deployment of games and simulations for purposes of education and training; human performance engineering; applications of games to health, public policy and strategic communication; game evaluation; serious game development.

Figure 2.6: Using serious games fields [12].

Among those areas, it is used in the field of training and education because it engages learners and brings better learning outcomes. Games can continuously motivate learners and challenge them to continue. This way, they keep being engaged with the subject material.

According to the Kipp-report [14], as shown in figure 2.7, compared with "teach others", the "Game based interactive learning" at the top position can produce the retention rate of 95% in terms of what learners remember after a period of time.

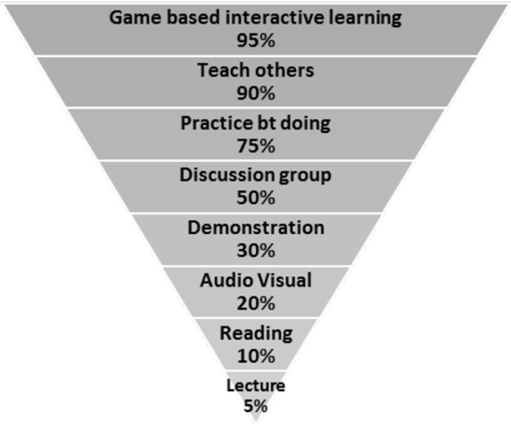


Figure 2.7: Game-based learning's rank in the learning pyramid [14].

This means that game-based learning(including serious games) can positively impact on the process of enhancing learning and memory, and establish a learning situation that attracts students' attention with challenge and entertainment [15]. Thus, the serious game also a good ideal for cybersecurity awareness training.

2.3.3 Cybersecurity Awareness Training

Serious games have received widespread attention in cybersecurity awareness training. One example is an online game Anti-Phishing Phil [16], created by Carnegie Mellon University. The game aims to teach users to keep good habits, avoid phishing fraud or attacks. In the game, the user plays a small fish role to determine whether the URL carried by the nearby worm is a phishing link. Through interactive operations, give users some hints to help them to distinguish whether the link is genuine or phishing. Although this game effectively helps people identify phishing websites, due to phishing websites' ever-changing nature, the examples given in the game cannot represent all the possible types of phishing websites, which has limitations for learning potential attacks.

Another example is the video game CyberCIEGE [17] developed by Naval Postgraduate School, a simulated network environment, allowing students to play different roles to understand network attacks. The game, including seven fundamental network attacks and customizable scenarios. Nevertheless, installing and configuring the game is troublesome and not very friendly to the user experience.

Internet Hero [18] is the learning game. Children need to solve four mini-games related to aspects of Internet use. In the game, players learn about the technical and social basis of using the Internet through their fictional world roles. A tutor role will be provided in the game to explain aspects of the Internet and games. Although the game was enjoyable for children, there are not many contents related to cybersecurity awareness training.

Moving forward, CSRAG [19] is a card game, aims to teach players the knowledge of software security and cybersecurity concepts while playing the game. This game has multiple features, such as role-playing, team-based learning, and extensive security content. However, this game requires participants to be in the physical world, it will take some time for novices to understand the rules of the game, and there are requirements for the number of participants in the game (games with less than two people will lose interest).

Finally, CybAR [20] was developed by Macquarie University, is applied to increase players' cybersecurity awareness and knowledge by the mobile augmented reality (AR) game. In particular, this game uses interactive features to show players the terrible consequences of not being aware of the potential cyber dangers. Train users through a series of game tasks applied quizzes learning principles to optimize the learning effect. This innovative pedagogical method can increase cybersecurity awareness after training, but this game only focuses on identifying potential attacks, thus ignoring the skills of educating users on how to deal with threats or management.

The table 2.2 on the following page shows the comparison between 5 serious games in cybersecurity awareness training and education.

Unlike the above five studies on the application of serious games in cybersecurity awareness training, our CyATP platform provides the serious game and exploratory interest learning pedagogic method to enhance their interest, learners can increase their security awareness knowledge and put it to use in their daily life.

Name	Type	Cybersecurity content	Purpose	Assessment method	Results	Features
Anti-Phishing Phil, 2007[16]	Web online game, 2D	Phishing attacks	To teach players how to identify phishing URLs, and use search engines to find legitimate sites.	Pre-Test and Post-Test, Usability Questionnaire	Improved learning and susceptibility of phishing	Easy to play Story line Time pressure
CyberCIEGE, 2014[17]	Video Games, 3D	Network security	Be responsible for achieving a balancing act minimizing the risk to the enterprise while allowing users to accomplish their goals.	Pre-Test and Post-Test, Usability Questionnaire	Generally positive	Play different roles Personalized learning Customize scenarios
Internet Hero, 2014[18]	Web online mini-games, 2D	Emails, Malicious programs, Social networks,	To solve different mini-games correlating to four aspects of Internet use, players to understand the basic technical or social aspects of these topics.	Data logging and Questionnaire	Children liked this game	Easy to play Story line Different types
CSRAG, 2019[19]	Card based game	Connection types Software system security	To make aware the technology players regarding security concepts, and the possible ways to identify potential threats in real environment.	Pre-Test and Post-Test, Usability Questionnaire	Positive effect on security learning outcomes	Play the role Content extensive Team-based learning
CybAR, 2020[20]	Mobile Augmented Reality Application, 3D	Incorporates all potential cyberattack techniques	To educate players about important cybersecurity related concepts.	Questionnaire	Improve cybersecurity awareness	Demonstrates the actual consequences of cybersecurity attacks Good interactivity Timely feedback

Table 2.2: Comparison of 5 serious games in cybersecurity awareness training and education.

Chapter 3

Training Content Generation

In this chapter, we will introduce the process of generating the training content. First, we introduce the overview of Machine Learning used to generate content. Then we introduce data preparation, feature engineering and training models respectively. Finally, we use the trained model to predict and generate training content.

3.1 Overview

There are different methods of Machine Learning, but the workflow for processing data is roughly the same. As shown in figure 3.1 , these processes including: Data collection, Data preparation, Choosing a model, Training, Evaluation, Parameter tuning and Prediction.

STEP 01
Data collection
STEP 02
Data preparation
STEP 03
Choosing a model
STEP 04
Training
STEP 05
Evaluation
STEP 06
Parameter tuning
STEP 07
Prediction

Figure 3.1: The steps of machine learning [21].

In this study, the overview of content generation is shown in figure 3.2 on the following page, which is roughly divided into three parts: Data preparation, Training, Development and Evaluation model, and Prediction. In section 3.2 data preparation, will discuss the sources of training data and predict data. In section 3.3, present how to process feature processing on the training data. In section 3.4, select the best model based on the results of evaluation and tuning parameters, and finally in section 3.5, use the trained model to predict and generate content.

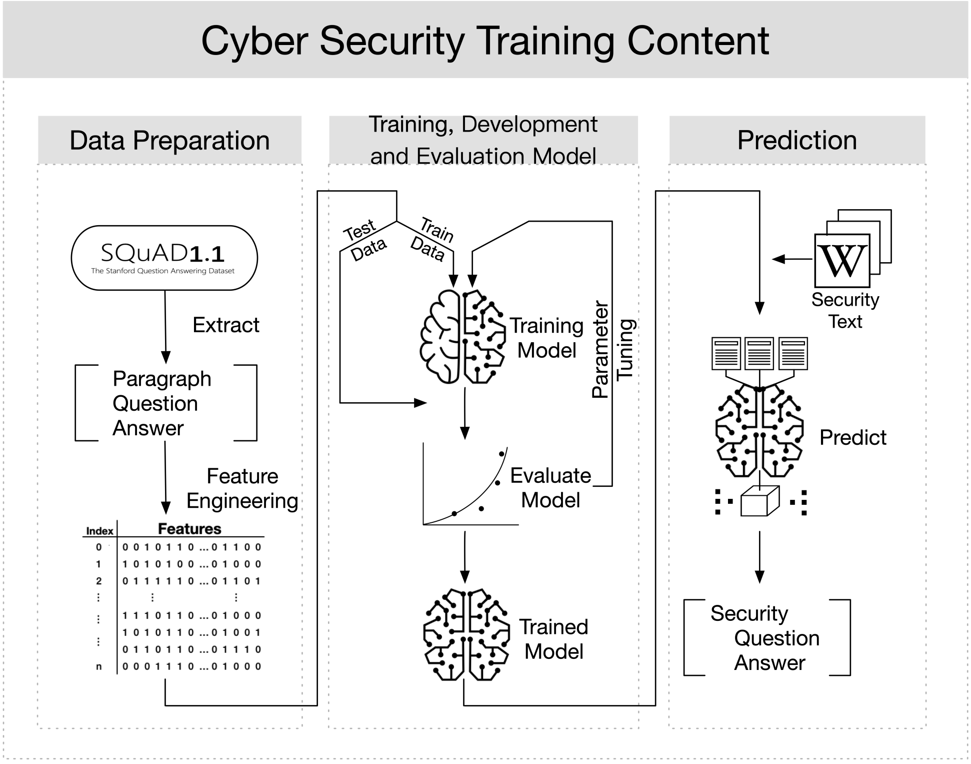


Figure 3.2: The overview of training content generation.

3.2 Data Preparation

There is a very famous quote in the Machine Learning field: data determines the upper bound of machine learning, the model and algorithm just approach this upper bound. This shows that data is a significant part of Machine Learning.

Usually, different datasets are used according to different research purposes. For example, use some representative public datasets. For image processing datasets: the handwritten digits dataset MNIST, the visual dataset ImageNet; For voice processing datasets: Free Spoken Digit, Free Music Archive (FMA); For text processing datasets: the news dataset Twenty Newsgroups, the sentiment analysis dataset Sentiment140.

However, for some specialized fields, it is not easy to find ready-made datasets. That needs researchers to collect and process the data by themselves. Data collection methods can use web crawlers, database pull, or API calls.

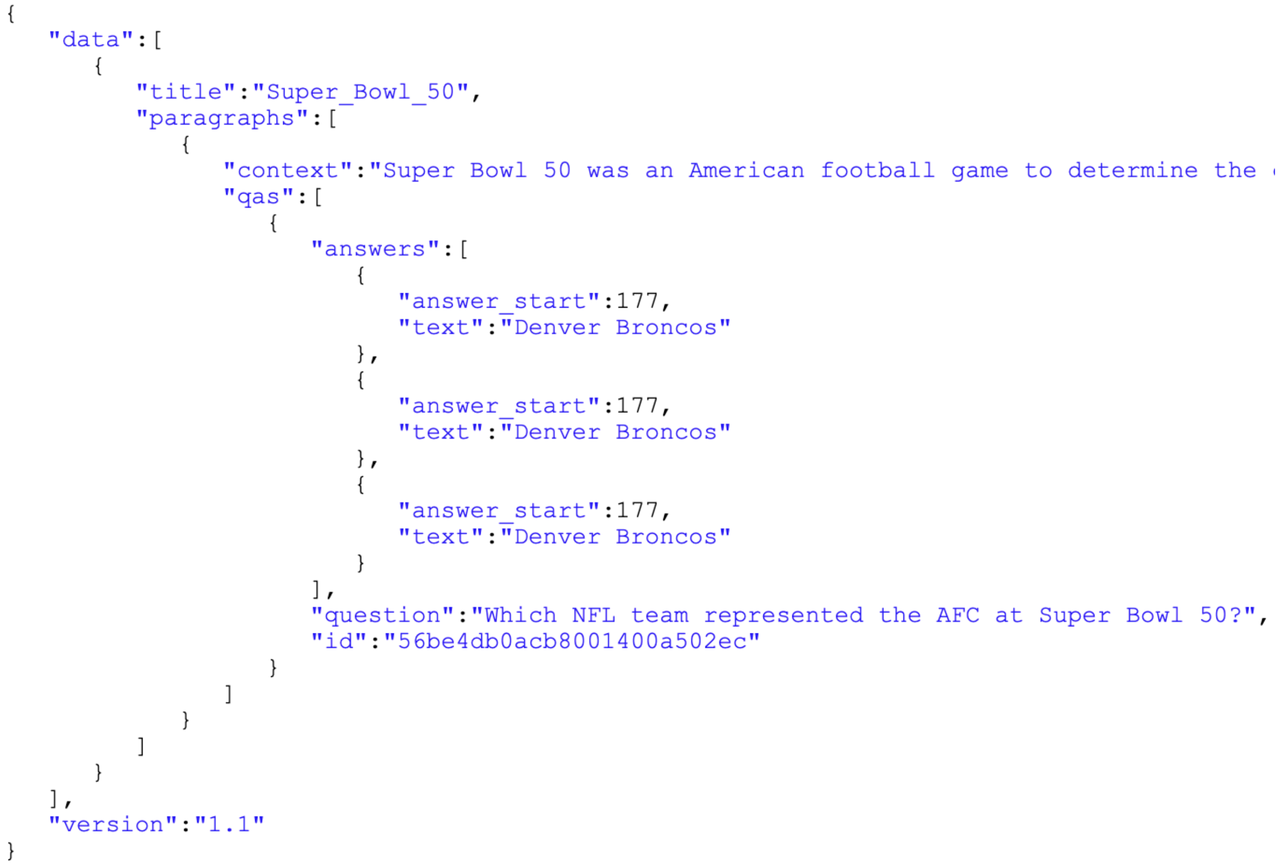
Compared with the dataset compiled by oneself, the public dataset is more representative, and data processing results are easier to be recognized. In addition, the public dataset can handle the problems of data overfitting, data deviation, and missing data better.

In this study, we use the public dataset for training. Because there is no suitable cybersecurtiy awareness training dataset, we collect it by ourselves as the predict dataset.

3.2.1 Training Dataset

We use the public dataset Stanford Question Answering Dataset (SQuAD1.1) [22] as training data. It includes 100,000+ question-answer pairs on more than 500 articles. The questions are asked by workers based on the paragraph of articles on Wikipedia. The answer to each question is a piece of text from the corresponding paragraph.

The original official data is divided into three parts: train sets, development sets, and test sets. Among them, train sets and development sets are available for download, and test sets are used for the official evaluation of machine learning models.



```
{
   "data":[
      {
         "title":"Super_Bowl_50",
         "paragraphs":[
            {
               "context":"Super Bowl 50 was an American football game to determine the
               "qas":[
                  {
                     "answers":[
                        {
                           "answer_start":177,
                           "text":"Denver Broncos"
                        },
                        {
                           "answer_start":177,
                           "text":"Denver Broncos"
                        },
                        {
                           "answer_start":177,
                           "text":"Denver Broncos"
                        }
                     ],
                     "question":"Which NFL team represented the AFC at Super Bowl 50?",
                     "id":"56be4db0acb8001400a502ec"
                  }
               ]
            }
         ]
      }
   ],
   "version":"1.1"
}
```

Figure 3.3: The example of SQuAD1.1 dataset.

As shown in figure 3.3, the structure of the dataset is divided into data and version. The data contains titles, paragraphs, context, qas, answer start location, answer text, question and id. In this research, we merge the official train sets and development sets as the training dataset, which has 490 articles, 20963 paragraphs, and 98169 questions. In all questions show in figure 3.4 on the following page, *what* type questions accounted for 58%, *who* type questions accounted for 10%, and *which* type questions accounted for 6.7%.

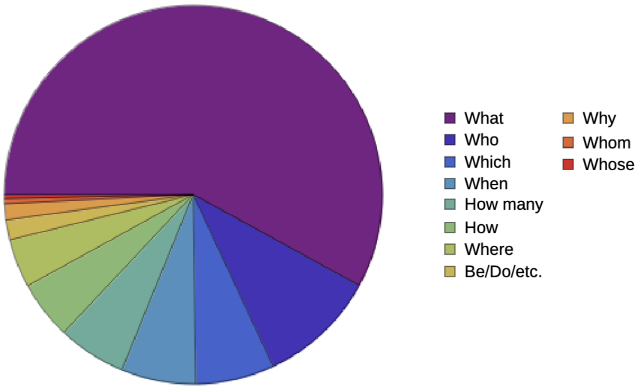


Figure 3.4: The distribution of question types in the SQuAD1.1 dataset.

3.2.2 Predict Dataset

Source of Keyword and Concept Map

Based on the research [23] of constructing the computer security concept map from the LOD database DBpedia, our research uses the keywords and concept maps part, to show the cybersecurity concept map to learners in a visual form, so that they can understand security knowledge more conveniently.

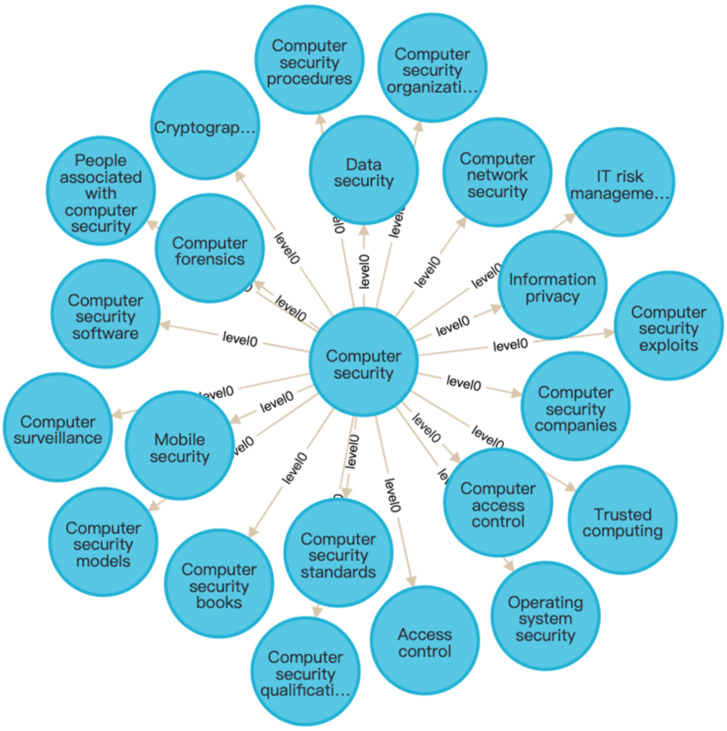

Computer security
Computer security procedures
Computer security organizati...
Cryptograp...
Data security
Computer network security
IT risk manageme...
People associated with computer security
Computer forensics
Information privacy
Computer security exploits
Computer security software
Computer security companies
Computer surveillance
Mobile security
Computer access control
Trusted computing
Computer security models
Computer security books
Computer security standards
Access control
Operating system security
Computer security qualificati...
level0

Figure 3.5: The keyword of "Computer security" and its concept map

As shown in figure 3.5, it is a concept map of the one level associated with "Computer security" as the root node. There are a total of 2640 keywords, and each word has its concept map.

Through the collected data keywords and concept maps, to provide users with more information security-related knowledge. In particular, the concept map method makes the knowledge nodes related, easy to establish a knowledge framework, and facilitates the use and memory of subsequent knowledge.

Source of Concept Text

In order to maintain the consistency of all data, the concept text is also from Wikipedia. Through the python library Wikipedia to get each keywords' concept text summary. While collecting the concept text data, use regular expressions to filter special characters in the text to ensure clean data. From table 3.1, we can see a total of 2640 keywords, of which 2315 keywords can query the concept text. Some keywords cannot find the concept text because of ambiguity errors. The same keyword has different conceptual interpretations under different fields, so ambiguity may arise when searching for keywords without specifying the division.

	Keyword
No Concept text	325
With Concept text	2315
Total	2640

Table 3.1: The information of concept map.

Organize all the predict data, the construction of the predict dataset can be see in table 3.2. There are a total of 8 levels of keywords and concept texts from level 0 to level 7. The corresponding number of each level can be viewed in the table below.

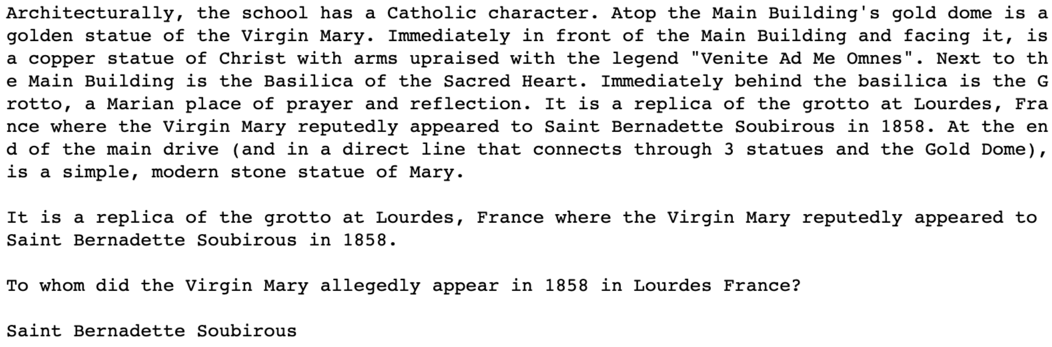
	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Total
Keyword	1	22	103	205	287	266	463	1293	2640
Concept text	1	17	90	182	266	234	402	1123	2315

Table 3.2: The information of predict dataset.

3.2.3 Data Preprocessing

To generate cloze questions and answers, we need to extract the paragraphs, sentences (containing the answer), questions and answers in SQuAD1.1. For example, figure 3.6 on the following page shows one piece of data, the first item is a paragraph, the second item is the sentence where the answer is, the third item is the question, and the fourth item is the answer.

Because the raw data cannot directly use, after extracting those data information, need to preprocess and standardize data to eliminate human error or incorrectness, and the impact of repeated data on the results.



```
Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a
golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is
a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to th
e Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the G
rotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, Fra
nce where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the en
d of the main drive (and in a direct line that connects through 3 statues and the Gold Dome),
is a simple, modern stone statue of Mary.

It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to
Saint Bernadette Soubirous in 1858.

To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?

Saint Bernadette Soubirous
```

Figure 3.6: The example of the extracted data.

1. Check out the missing data. Before using the dataset, we first check the completeness of the data and whether it lacks attribute values. Focus on whether the questions in the data set have corresponding answers. If the answers are missing, the corresponding questions will not be used. Of course, if any of the paragraph, question and answer is missing in a piece of data, this data will be deleted.

2. Delete inappropriate data. We expect that all questions and answers are based on paragraphs. Check the similarity of the question and sentence, the answer and sentence in each piece of data. Delete the data with the similarity of 0.

3. Clean data. There are some duplicate values and inconsistencies in the data. Delete these data so as not to affect the accuracy of the results.

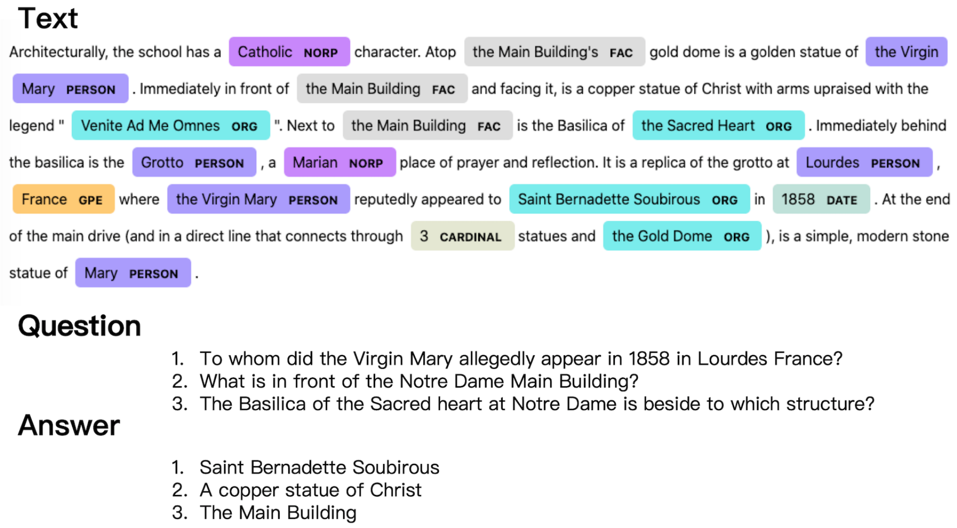
3.2.4 Feature Engineering

Feature engineering uses the prepared data in Section 3.2.3, to create the features expected by the Machine Learning Model. This process is to convert text into data that can be used for Machine Learning.

Features Selection

The goal is to generate text-based cloze questions and answers. First, we need to understand the relationship between the questions-answers and the text. An example from the prepared data in figure 3.7 on the next page, that shows people always like to ask W-questions (what, when, where, who, why), most of these objects are entities or noun chunks (marked by color).

Through this habit of people, we selected 11 items to construct features data. There are "Is_Answer" (Does the word or noun chunk appear in the answer), "TitleId" (The text id of the word or noun chunk), "ParagraphId" (The paragraph id of the word or noun chunk), "SentenceId" (The sentence id of the word or noun chunk), "InSentencePosition" (The position of the word in the sentence), "Word_Count" (the count of word or noun chunk), "NER" (The named entity recognition), "POS" (The simple part-of-speech tag), "TAG" (The detailed part-of speech tag), "DEP"(Syntactic dependency),"Is_Alpha" (Is



Text

Architecturally, the school has a Catholic NORP character. Atop the Main Building's FAC gold dome is a golden statue of the Virgin Mary PERSON . Immediately in front of the Main Building FAC and facing it, is a copper statue of Christ with arms upraised with the legend " Venite Ad Me Omnes ORG ". Next to the Main Building FAC is the Basilica of the Sacred Heart ORG . Immediately behind the basilica is the Grotto PERSON , a Marian NORP place of prayer and reflection. It is a replica of the grotto at Lourdes PERSON , France GPE where the Virgin Mary PERSON reputedly appeared to Saint Bernadette Soubirous ORG in 1858 DATE . At the end of the main drive (and in a direct line that connects through 3 CARDINAL statues and the Gold Dome ORG), is a simple, modern stone statue of Mary PERSON .

Question

1. To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?
2. What is in front of the Notre Dame Main Building?
3. The Basilica of the Sacred heart at Notre Dame is beside to which structure?

Answer

1. Saint Bernadette Soubirous
2. A copper statue of Christ
3. The Main Building

Figure 3.7: The example of the prepared data.

the word or noun chunk an alpha character), "Is_Stop" (Is the word or noun chunk a stop word). Among them, "NER", "POS", "Tag", "Dep", "Is_alpha", and "Is_stop" items are realized through Spacy [24] python library. The types of NER can be seen in figure 3.8. Examples of POS can be seen in figure 3.9 on the next page. The details of TAG can be seen in figure 3.10 on the following page. The example labels of DEP can be seen in figure 3.11 on page 26.

Type	Description
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
...	...

Figure 3.8: The samples of NER (Named Entity Recognition).

POS	Description	Examples
ADJ	adjective	big, old, green, incomprehensible, first
ADP	adposition	in, to, during
ADV	adverb	very, tomorrow, down, where, there
AUX	auxiliary	is, has (done), will (do), should (do)
CONJ	conjunction	and, or, but
CCONJ	coordinating conjunction	and, or, but
DET	determiner	a, an, the
INTJ	interjection	psst, ouch, bravo, hello
NOUN	noun	girl, cat, tree, air, beauty
...	...	...

Figure 3.9: The samples of POS (Part-of-Speech Tag).

TAG	POS	Morphology	Description
NN	NOUN	Number=sing	noun, singular or mass

NNP	PROPN	NounType=prop
		Number=sing
			noun, proper singular

NNPS	PROPN	NounType=prop
		Number=plur
			noun, proper plural

NNS	NOUN	Number=plur	noun, plural
PDT	DET		predeterminer
POS	PART	Poss=yes	possessive ending
PRP	PRON	PronType=prs	pronoun, personal

PRP$	DET	PronType=prs
		Poss=yes
			pronoun, possessive

RB	ADV	Degree=pos	adverb
...	...	...	...

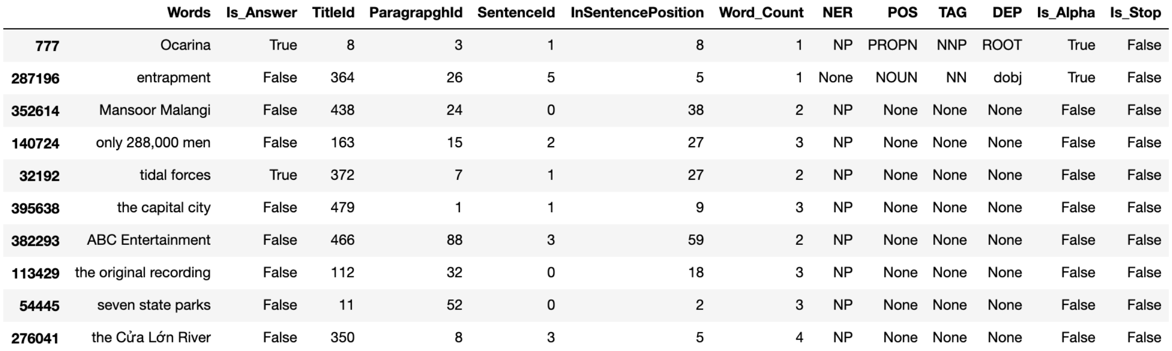
Figure 3.10: The detailed of TAG (Detailed Part-of-Speech Tag).

Label	Description
acl	clausal modifier of noun (adjectival clause)
advcl	adverbial clause modifier
advmod	adverbial modifier
amod	adjectival modifier
appos	appositional modifier
aux	auxiliary
case	case marking
cc	coordinating conjunction
ccomp	clausal complement
...	...

Figure 3.11: The detail labels of DEP (Syntactic Dependency).

Extract features

Extract each noun and noun chunk from the prepared set, to determine whether they have the selected features. Due to the extraction of features from words and noun chunks, there will be duplicate data. It is necessary to deduplicate and delete strange data after extracting features. As shown in figure 3.12, it is ten random examples after extracting features.

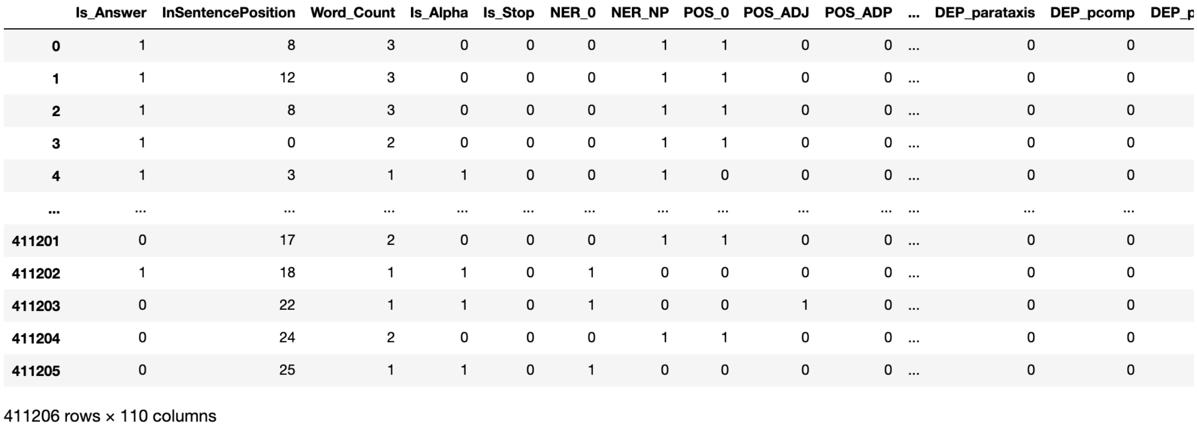


	Words	Is_Answer	TitleId	ParagrapghId	SentenceId	InSentencePosition	Word_Count	NER	POS	TAG	DEP	Is_Alpha	Is_Stop
777	Ocarina	True	8	3	1	8	1	NP	PROPN	NNP	ROOT	True	False
287196	entrapment	False	364	26	5	5	1	None	NOUN	NN	dobj	True	False
352614	Mansoor Malangi	False	438	24	0	38	2	NP	None	None	None	False	False
140724	only 288,000 men	False	163	15	2	27	3	NP	None	None	None	False	False
32192	tidal forces	True	372	7	1	27	2	NP	None	None	None	False	False
395638	the capital city	False	479	1	1	9	3	NP	None	None	None	False	False
382293	ABC Entertainment	False	466	88	3	59	2	NP	None	None	None	False	False
113429	the original recording	False	112	32	0	18	3	NP	None	None	None	False	False
54445	seven state parks	False	11	52	0	2	3	NP	None	None	None	False	False
276041	the Cửa Lớn River	False	350	8	3	5	4	NP	None	None	None	False	False

Figure 3.12: Example data after extracting features.

Encoding

Although the data after the feature extraction has a good structure, it cannot be directly put into the model, because the model cannot recognize the string, and it needs to be converted into a number before it can be put into the model. Use One-Hot Encoding to digitize discontinuous and discrete features. The specific process of One-Hot encoding mainly uses N-bit status registers to encode N states. Each state has its independent register bit, and only one bit is valid at any time. It can avoid the problem that Integer Encoding allows the model to assume the natural ordering between categories and cause poor performance. As shown in figure 3.13, the example of data after encoding, the "Words" item is deleted, and the original 11 features items are expanded to 118.



	Is_Answer	InSentencePosition	Word_Count	Is_Alpha	Is_Stop	NER_0	NER_NP	POS_0	POS_ADJ	POS_ADP	...	DEP_parataxis	DEP_pcomp	DEP_p
0	1	8	3	0	0	0	1	1	0	0	...	0	0	
1	1	12	3	0	0	0	1	1	0	0	...	0	0	
2	1	8	3	0	0	0	1	1	0	0	...	0	0	
3	1	0	2	0	0	0	1	1	0	0	...	0	0	
4	1	3	1	1	0	0	1	0	0	0	...	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
411201	0	17	2	0	0	0	1	1	0	0	...	0	0	
411202	1	18	1	1	0	1	0	0	0	0	...	0	0	
411203	0	22	1	1	0	1	0	0	1	0	...	0	0	
411204	0	24	2	0	0	0	1	1	0	0	...	0	0	
411205	0	25	1	1	0	1	0	0	0	0	...	0	0	

411206 rows × 110 columns

Figure 3.13: Example data after encoding.

3.2.5 Dealing with Imbalanced Data

After the encoding is completed, the sample data that is not an answer is much larger than the sample data that is an answer, and this situation is imbalanced data. If this problem is not dealt with, the model will always predict the side with more data and lose the meaning of learning. We use the following four methods to balance the minority samples and select the most suitable method according to the training results in section 3.3.3.

Resampling adjustment method

For solving imbalanced data, oversampling and undersampling are the two most basic methods.

- Oversampling is a method of randomly sampling from minority samples to increase the size of minority samples. However, some samples will repeatedly appear in the dataset after oversampling, which may lead to the overfit of the trained model.

- Undersampling is to randomly select a small number of samples from the majority of samples to balance the proportion of minority samples. However, in order to ensure the balance of the dataset, the undersampling method discards some data, which may be particularly important for the training of the model, so that the model only learns a part of the data.

Generate synthetic samples

The basic idea of SMOTE (Synthetic Minority Oversampling Technique) algorithm and ADASYN (Adaptive Synthetic) algorithm is to analyze minority samples, artificially synthesize new samples based on minority samples and add them to the dataset [25].

The algorithm flow is as follows:

1. Randomly select a minority sample X_i
2. Find the k neareast-neighbors closest to X_i (as shown in figure 3.14 the blue circle with 3 nearest-neighbors)
3. Randomly select a nearest neighbor X_{zi}
4. Generate X_{new} according to the $X_{new} = X_i + \lambda \times (X_{zi} - X_i)$, λ is a random number in the range [0,1].

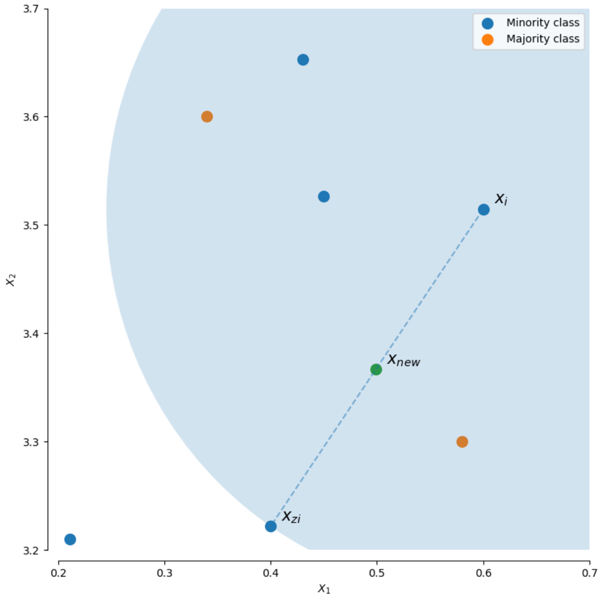


Figure 3.14: The sample generation in the synthetic algorithm [25].

The difference between SMOTE and ADASYN algorithm is to select X_i before generating new samples. Although the use of synthetic algorithms can eliminate imbalance and improve the efficiency of learning [26], the selected minority class samples are surrounded by the majority class samples, then this type of sample may be noise, and the newly samples will overlap most of the surrounding majority samples, making it difficult to classify.

Comparison

10000 sample datas are generated in figure 3.15, and each class accounts for 94% (class 2: 9345), 5% (class 1: 523) and 1% (class 0: 132) respectively. Using these samples we compare the 4 different methods to deal with imbalanced data.

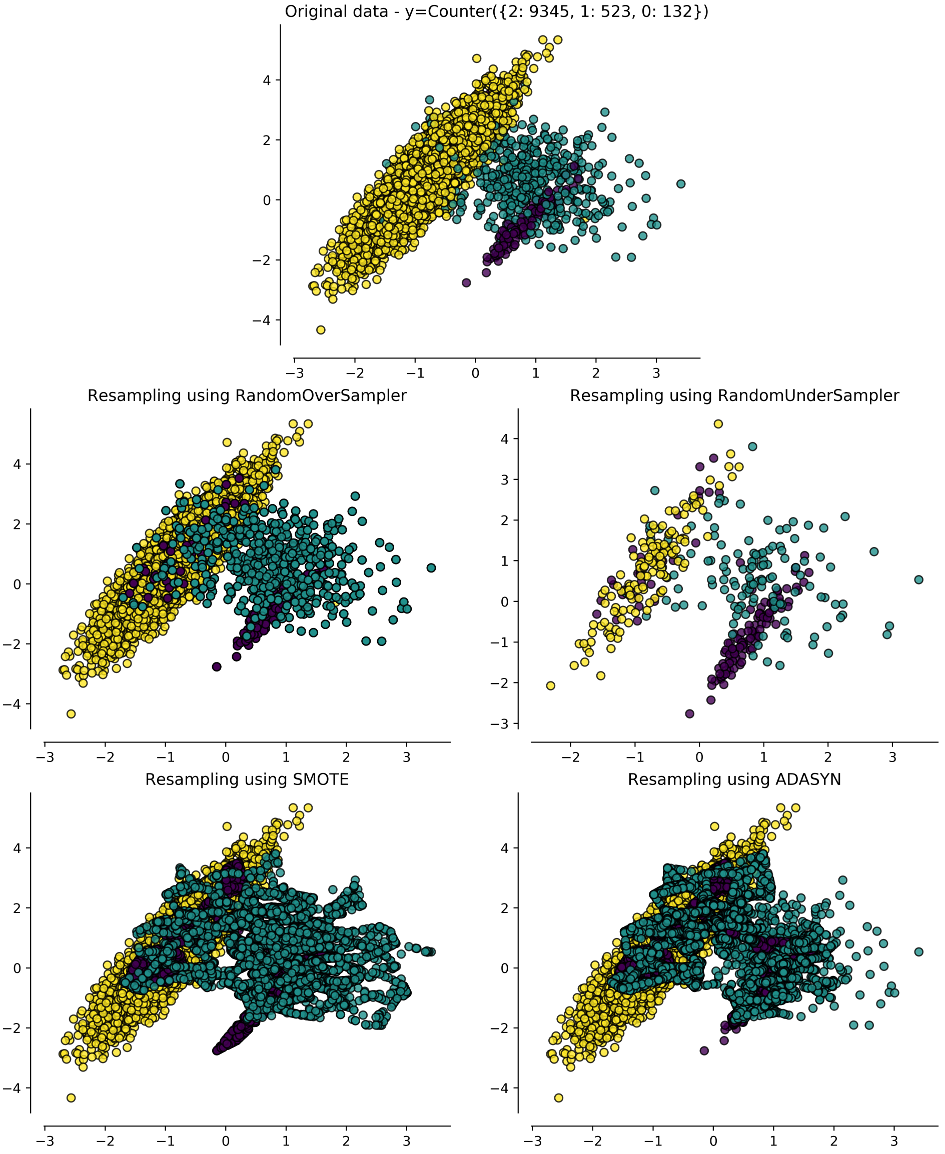


Figure 3.15: Compare 4 methods to processing imbalanced data.

3.2.6 Training Data and Test Data

The data after feature engineering is divided into 80% training data and 20% test data. The training data is used to train the Naive Bayes Models, and the test data is used to provide an unbiased evaluation of the trained model.

The training data is then processed according to the 4 methods introduced in Section 3.2.5, and finally the data of the training data is shown in table 3.3, and the data of the test data is shown in table 3.4.

Number of Data	Original Data	Resampling Techniques		Generate Synthetic Samples	
		Oversample	Undersample	SMOTE	ADASYN
Total	328964	543858	114070	543858	530960
False Label	271929	271929	57035	271929	271929
True Label	57035	271929	57035	271929	259031

Table 3.3: The information of training data.

Total Data	82242
False Label	67888
True Label	14354

Table 3.4: The information of test data.

3.3 Training Models

Put 5 types of training data (Original Data, Oversampling Data, Undersampling Data, SMOTE Data, ADASYN Data) into 4 types of Naive Bayes model (Gaussian Naive Bayes, Bernoulli Naive Bayes, Multinomial Naive Bayes, Complement Naive Bayes) for training. According to the evaluation criteria, tune the parameters and select the best model.

3.3.1 Training and Evaluation Model

In this research, we use the computing server Cray XC40 to train the models. Each training job used 4 nodes to improve the training speed. The size of the node is Intel Xeon E5-2695v4 (2.1GHz, 18Core x 2), 128GB Memory (16GB DDR4-2133 x8).

The trained results are shown in table 3.5 on the following page.

We used 9 evaluation metrics to quantify the performance of the models. The Accuracy, Precision, Recall, F1 score are calculated by the confusion matrix. Each column of the confusion matrix represents the predicted class, and the total number of each column represents the number of data predicted to be that class; each row represents the true attribution class of the data, and the total number of data in each row represents the number of data instances of that class. An example of Bernoulli Naive Bayes trained results is shown in figure 3.16 on page 32, TN means true negative, predicted to be false label, actual label is also false; TP means true positive, predicted to be true label, actual label is also true; FP means false positive, predicted to be true label, actual label is false; FN means false negative, predicted to be false label, actual label is true.

Models	Time Cost	Train Score	Accuracy Score	Precision Score	Recall Score	F1 Score	Logloss Score	Brier Score	AUC
GaussianNB_original	00:01:496647	0.21188945	0.21206926	0.17929026	0.98237425	0.30323753	27.0077149	0.78793511	0.52564678
GaussianNB_oversampled	00:02:245883	0.5163425	0.21197199	0.17927203	0.98237425	0.30321145	27.0674237	0.78802325	0.52500933
GaussianNB_downsampled	00:00:385076	0.52438853	0.22906787	0.1821728	0.9793089	0.30719976	25.8808404	0.7732826	0.53463587
GaussianNB_smote	00:02:456814	0.51891119	0.21230028	0.17926832	0.98181692	0.30317959	27.1991909	0.78769976	0.51592697
GaussianNB_adasyn	00:01:969841	0.50766159	0.21222733	0.17913221	0.98077191	0.3029351	27.1925823	0.78777248	0.51612131
MultinomialNB_original	00:00:401142	0.84517759	0.8437538	0.91870824	0.11495054	0.20433437	0.43041942	0.13152304	0.6396292
MultinomialNB_oversampled	00:00:665214	0.56244645	0.74626103	0.27114952	0.26884492	0.2699923	0.63081726	0.22125465	0.63827303
MultinomialNB_downsampled	00:00:140885	0.5587271	0.75863914	0.28065134	0.24494914	0.26158768	0.62945755	0.22066417	0.64000931
MultinomialNB_smote	00:00:648798	0.57181102	0.71930401	0.26296357	0.33739724	0.2955662	0.63157566	0.22176806	0.63741558
MultinomialNB_adasyn	00:00:719254	0.52680428	0.32437198	0.18890315	0.8716734	0.31051384	0.68917234	0.25140331	0.64300326
ComplementNB_original	00:00:411723	0.7516628	0.74843754	0.27389535	0.26731225	0.27056376	0.63042638	0.22117735	0.63962916
ComplementNB_oversampled	00:00:591467	0.56244645	0.74626103	0.27114952	0.26884492	0.2699923	0.63081726	0.22125465	0.63827299
ComplementNB_downsampled	00:00:114780	0.5587271	0.75863914	0.28065134	0.24494914	0.26158768	0.62945755	0.22066417	0.64000927
ComplementNB_smote	00:00:785985	0.57181102	0.71930401	0.26296357	0.33739724	0.2955662	0.63157566	0.22176806	0.63741558
ComplementNB_adasyn	00:00:711147	0.52609424	0.32248729	0.1885747	0.87250941	0.31012282	0.70493198	0.25913893	0.64300326
BernoulliNB_original	00:00:763727	0.84551805	0.84410642	0.92324682	0.11648321	0.20686669	0.43136699	0.13120632	0.59650404
BernoulliNB_oversampled	00:01:402953	0.55536188	0.8035432	0.36392453	0.16796712	0.2298489	0.62748113	0.22051391	0.59643657
BernoulliNB_downsampled	00:00:264403	0.55451915	0.80326354	0.36287173	0.16831545	0.22996383	0.62580716	0.21972351	0.59515103
BernoulliNB_smote	00:01:352380	0.55599991	0.80422412	0.3662942	0.16671311	0.22913774	0.62906228	0.22132073	0.59639772
BernoulliNB_adasyn	00:01:248105	0.52538044	0.32175774	0.18776568	0.86777205	0.30872949	0.76578037	0.28796226	0.57645776

Table 3.5: The trained results

The matrix shows that 14354 data are labeled as answer, 67888 data are labeled as not answer. The model predicted that 6532 data as answer and 75710 data are not answer. Use the following metrics to evaluate the performance of the model:

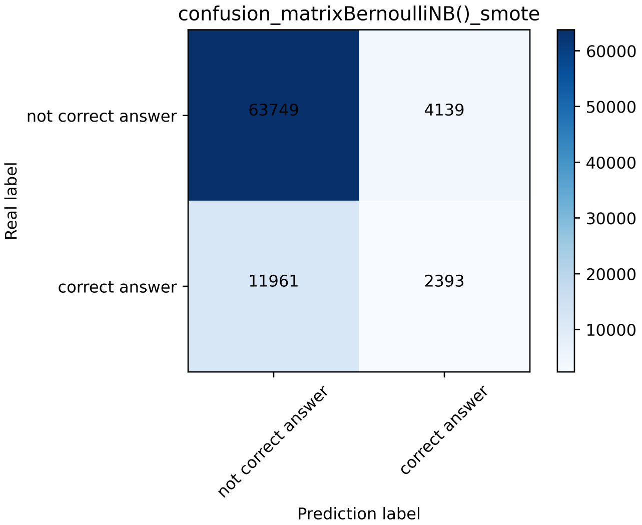


Figure 3.16: The confusion matrix of BernoulliNB model.

- Time cost. Time spent training the model.
- Train Score. The accuracy of the model on the training data.
- Accuracy Score. The accuracy of the model on the test data.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision Score. The correct rate in the predicted answer.

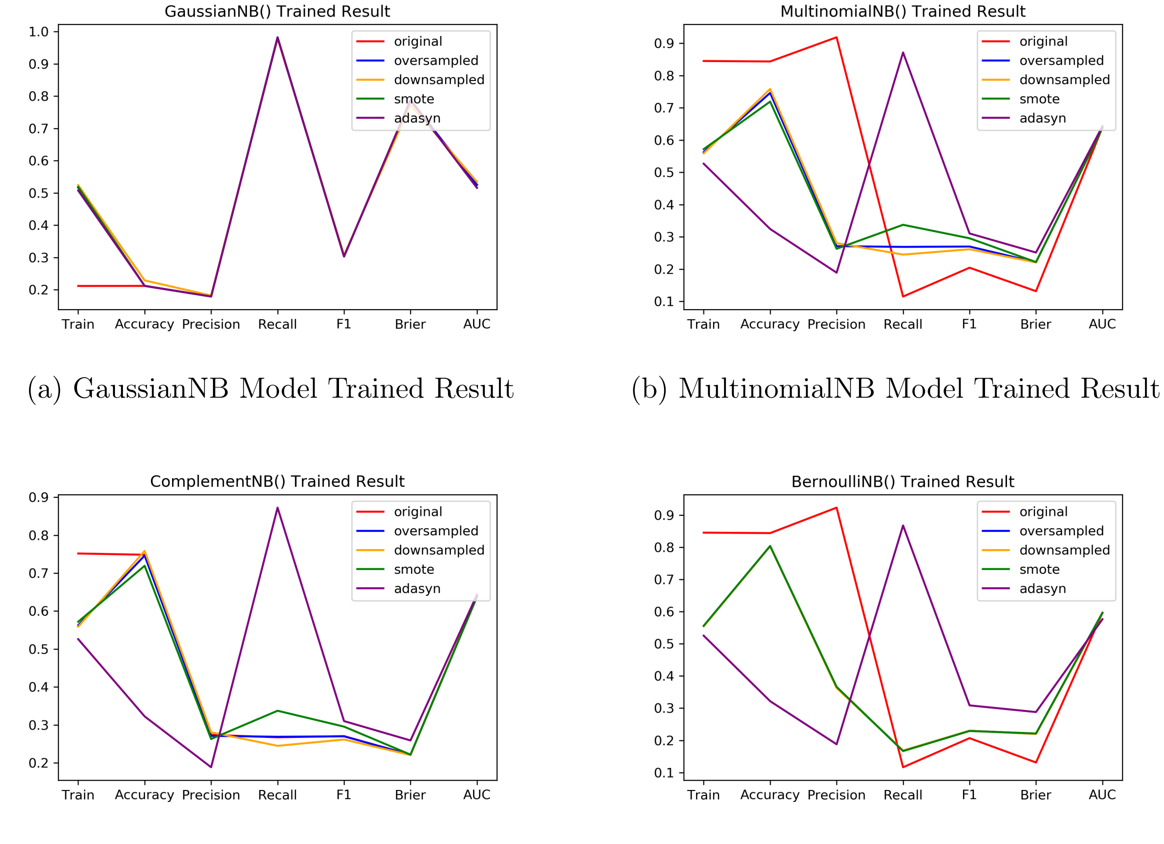
$$Precision = \frac{TP}{TP + FP}$$

- Recall Score. The proportion of predicted answer in actual answers.

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

- F1 Score. The weighted average between precision and recall. It is useful when dealing with unbalanced samples.

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



(a) GaussianNB Model Trained Result

(b) MultinomialNB Model Trained Result

(c) ComplementNB Model Trained Result

(d) BernoulliNB Model Trained Result

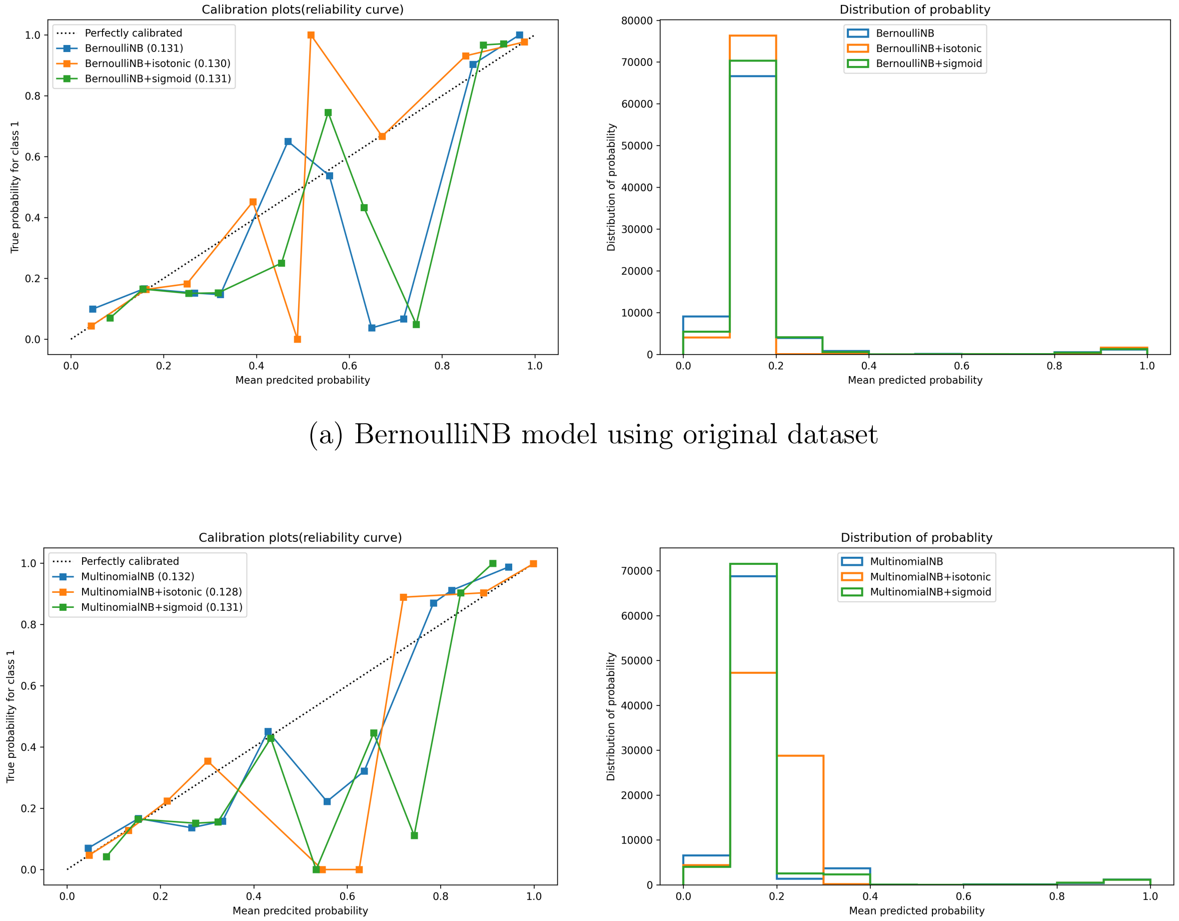
Figure 3.17: The trained results of 4 Naive Bayes models.

- Logloss Score. That is used to measure the degree of inconsistency between the predicted value of the model and the true value. It is a non-negative real-valued function. The smaller the loss function, the higher the accuracy of the prediction.

- Brier Score. It is only used to evaluate two classification problems, measuring the error between the probability of the class predicted by the model and the true value. The lower the value of this score, the better the prediction. The score for a perfect prediction is 0. The worst score is 1. If the Brier score hovers around 0.5 points, it is difficult to determine the quality of the model.

- AUC. The ROC (Receiver Operating Characteristic) curve is the relationship curve between FP and TP, and AUC (Area Under the Curve) is the area under the ROC curve. The larger the area, the better the performance of the model.

As figure 3.17a shows the trained results of GaussionNB model, roughly the same results for different data sets, with low accuracy, large errors, and high Logloss score. The model tends to divide the data into true class. From the AUC score, the model basically does not have the ability to classification. So this model is not suitable for this research.



(a) BernoulliNB model using original dataset

(b) MultinomialNB model using original dataset

Figure 3.18: Reliability curve and probability distribution of the BernoulliNB and MultinomialNB model.

The figure 3.17b on the previous page shows MultinomialNB trained results, the model has a relatively high accuracy on the test data. Since the training data is processed with balanced data, the scores of training accuracy and test accuracy will be different. The Logloss score is small, model trends classify the data as false label, and the AUC score performs better.

The ComplementNB results as figure 3.17c on the preceding page shows, although the model is nice, it did not perform so well in accuracy compared to the MultinomialNB and BernoulliNB. And the model tends to predict the data as false label.

The figure 3.17d on the previous page shows BernoulliNB trained results, the accuracy of predicting the test data can reach up to 0.8441, with small errors. The model tends to predict the data as false label. The AUC score is relatively high.

In section 3.2.5, we discussed the use of ADASYN to deal with imbalanced data. As show in table 3.5 on page 31, the performance of the four models on this data is not good, no matter the accuracy of the training data and the test data is very low, this way of processing data is not suitable for this study.

From the probability distribution histogram on the figure 3.18 on the previous page, when using the original data, the model divides most of the words into not answer. In this way, the model's insufficient learning for the minority classes will affect the prediction results, this way of processing data also not suitable for this study.

Through the analysis of the trained results, we retained two models (MultinomialNB and BernoulliNB) and three data processing method (oversampling, undersampling and SMOTE).

3.3.2 Parameter Tuning

There are two calibration methods for probabilistic models: the parametric approach of sigmoid scaling [27] and non-parametric approach of isotonic regression [28]. Sigmoid scaling is simpler and is suitable for reliability diagrams with the S-shape. Isotonic regression is more complex, requires a lot more data (otherwise it may overfit), but can support reliability diagrams with different shapes (is nonparametric). Sigmod scaling is most effective when the distortion in the predicted probabilities is sigmoid-shaped. Isotonic regression is a more powerful calibration method that can correct any monotonic distortion. Unfortunately, this extra power comes at a price. A learning curve analysis shows that isotonic regression is more prone to overfitting, and thus performs worse than sigmod scaling, when data is scarce.

The calibration results of the model and the method of processing data after the screening in Section 3.3.1 are shown in table 3.6 on the following page.

From the training result data, the BernoulliNB model has been calibrated by isotonic to improve its accuracy, reaching the score of 0.84. This result is the same as the conclusion of Niculescu-Mizil paper [29]. When there is enough data to train the model, the calibration of isotonic regression is more powerful and can also prevent overfitting.

3.3.3 Select Model

Compare with other models, the BernoulliNB model after isotonic calibration has high accuracy, low logloss score and less error. Since the scores of the previous 9 evaluation metrics are similar, the reliability curve and the probability distribution histogram are used to select the best model.

The calibration curve (reliability curve) uses the bucket method to discretize the continuous data and visually observe whether the predicted probability of the classification model is close to the true probability. The perfectly calibration curve is the diagonal line, that means the predicted value is exactly the same as the true value.

Models	Time Cost	Train Score	Accuracy Score	Precision Score	Recall Score	F1 Score	Logloss Score	Brier Score	AUC
MultinomialNB_oversample	00:00:576803	0.562446	0.746261	0.27115	0.268845	0.269992	0.630817	0.221255	0.638273
MultinomialNB+isotonic_oversample	00:01:463865	0.609906	0.615525	0.249463	0.598857	0.352208	0.625546	0.2191	0.648203
MultinomialNB+sigmoid_oversample	00:03:055366	0.591044	0.672369	0.256262	0.461126	0.329443	0.648766	0.229372	0.638275
MultinomialNB_downsample	00:00:127357	0.558727	0.758639	0.280651	0.244949	0.261588	0.629458	0.220664	0.640009
MultinomialNB+isotonic_downsample	00:00:305950	0.607942	0.616534	0.249905	0.598091	0.352516	0.623782	0.218352	0.651235
MultinomialNB+sigmoid_downsample	00:00:438143	0.594863	0.673804	0.26042	0.472273	0.335719	0.64773	0.228882	0.639889
MultinomialNB_smote	00:00:586870	0.569628	0.731524	0.267065	0.308555	0.286315	0.631791	0.221792	0.639088
MultinomialNB+isotonic_smote	00:01:535564	0.610014	0.609324	0.247371	0.606312	0.351381	0.624913	0.21924	0.648463
MultinomialNB+sigmoid_smote	00:03:798787	0.599515	0.660113	0.256857	0.500418	0.339469	0.647474	0.229004	0.63911
BernoulliNB_oversample	00:01:738604	0.555362	0.803543	0.363925	0.167967	0.229849	0.627481	0.220514	0.596431
BernoulliNB+isotonic_oversample	00:02:507909	0.559462	0.842307	0.833414	0.120594	0.210699	0.63353	0.222883	0.600809
BernoulliNB+sigmoid_oversample	00:04:194564	0.557149	0.346988	0.194541	0.872997	0.318179	0.650714	0.230959	0.596403
BernoulliNB_downsample	00:00:357346	0.554519	0.803264	0.362872	0.168315	0.229964	0.625807	0.219724	0.595156
BernoulliNB+isotonic_downsample	00:00:518870	0.559595	0.843632	0.888254	0.119061	0.209977	0.632775	0.222656	0.600747
BernoulliNB+sigmoid_downsample	00:00:646302	0.553756	0.80133	0.354706	0.168803	0.228747	0.649988	0.23054	0.595022
BernoulliNB_smote	00:01:754716	0.556325	0.804321	0.366825	0.166852	0.229373	0.628889	0.221135	0.596527
BernoulliNB+isotonic_smote	00:02:558848	0.560731	0.842562	0.845285	0.119897	0.210006	0.632601	0.222858	0.600737
BernoulliNB+sigmoid_smote	00:04:810937	0.558269	0.34801	0.194606	0.871604	0.318172	0.649835	0.230737	0.596456

Table 3.6: The trained results of MultinomialNB and BernoulliNB after parameter tuning

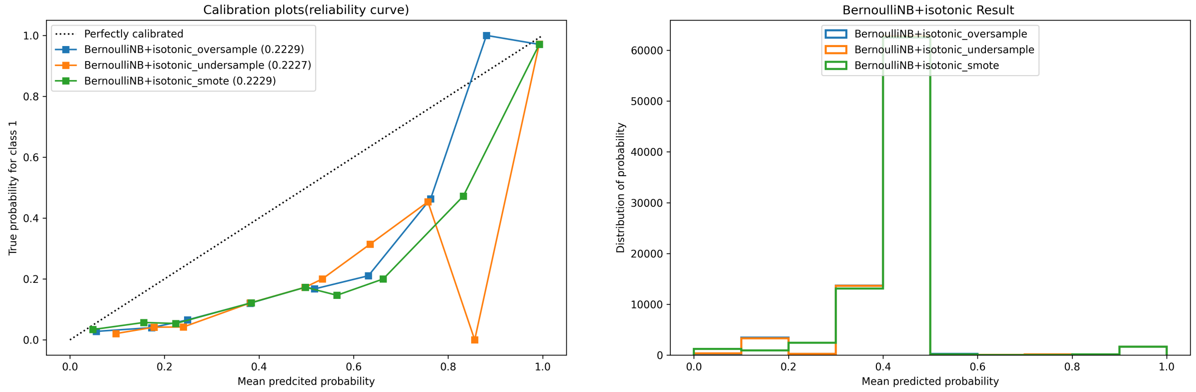


Figure 3.19: Reliability curve and probability distribution of the BernoulliNB model after isotonic calibration

As figure 3.19 shows the calibration plots and probability distribution of BernoulliNB model after isotonic calibration. The probability histogram distributions of the three data are almost the same, but the reliability curves are very different.

The orange one used the undersampling data, and there is an abnormal value between 0.8 and 1.0 on the abscissa, which makes the curve tortuous and deviate from the perfectly calibrated line.

The blue one used the oversampling data, the curve is roughly s-shaped. Between 0.0 and 0.9 on the abscissa, the curve is located below the diagonal, that means the model overestimates the low probability. Between 0.9 and 1.0, the curve is located above the diagonal, the model underestimates the high probability.

The green one used the SMOTE data. Although the curve is not closest to the diagonal, it is generally smooth and stable. The lower Brier score indicates that the prediction error is small.

After analyzing the training results, this research uses the data processed by the SMOTE method to train the BernoulliNB model after the isotonic calibration.

3.4 Predict and Generate Content

Use the predict data prepared in Section 3.2 and the trained model in Section 3.3 to make prediction and generation. The process is as follows:

1. The predict data is also subjected to preprocessing and feature engineering, and transformed into the data form that the model can use.

2. Put the data into the trained model for prediction, generate cloze question and answer pairs.

3. According to the predicted answer, use Gensim python library to generate other similarity options.

4. Select the generated questions with less than 25 words as crossword puzzles' clues.

5. Combine all predicted and collected data, store it in JSON file format.

In step 4, we selected the questions with less than 25 words. Because one of the crossword puzzles' characteristics is that the clues are as short as possible to reflect the answer to the greatest extent. Some research shows that the sentence longer than 25 words aren't accessible [30]. And in the short-term memory, as many as 20-25 words can be memorized in groups [31]. So based on the number of words in the sentence acceptable to people, the clues were selected.

The structure of the cybersecurity learning content dataset is shown in figure 3.20 on the following page, including "Keywords", "Level", "Text", "Questions", "Question count" and "Puzzle". In Section 4, that will be used to build the database and provide users with learning materials.

```
[
   {
      "Keyword":"Computer security",
      "Level":0,
      "Text":"Computer security, cybersecurity or information technology security  is the protection of computer
      "Questions":[
         {
            "question":"_____  is the protection of computer systems and networks from the theft of or damage to
            "answer":"computer security",
            "raw_choices":[
               "infosec",
               "cryptography",
               "encryption"
            ],
            "decided_choices":[
               "encryption",
               "cryptography",
               "computer security",
               "infosec"
            ]
         },
         {
            "question":" _____ is the protection of computer systems and networks from the theft of or damage to
            "answer":"cybersecurity",
            "raw_choices":[
               "cryptography",
               "counterterror",
               "dhs"
            ],
            "decided_choices":[
               "cryptography",
               "counterterror",
               "cybersecurity",
               "dhs"
            ]
         },
         {
            "question":"The field is becoming more important due to increased reliance on computer systems, the
            "answer":"smartphones",
            "raw_choices":[
               "handset",
               "netbook",
               "android",
               "nokia",
               "palmtop",
               "ios"
            ],
            "decided_choices":[
               "ios",
               "smartphones",
               "nokia",
               "android"
            ]
         },
        ...
      ],
      "Question_Count":4,
      "Puzzle":[
         {
            "answer":"cybersecurity",
            "clue":"Owing to its complexity, both in terms of politics and technology, _____ is also one of the
         }
      ]
   },
   ...
]
```

Figure 3.20: The structure of the cybersecurity learning content dataset.

Chapter 4

Cybersecurity Awareness Training Platform

This chapter, we talk about using the generated and collected data in Chapter 3 to build a cybersecurity content database. we build a web application using this database to provide learners with a platform for cybersecurity awareness training. First, we introduce the platform's framework, then talk about the function and usage of each page separately, and finally explain the implementation method.

4.1 Framework

CyATP (Cybersecurity Awareness Training Platform) is a web platform for cybersecurity awareness training that makes use of Natural Language Generation (NLG) techniques to automatically generate the training content; the serious game approach is employed for learning purposes. Using this platform, learners can increase their security awareness knowledge and use it in their daily lives.

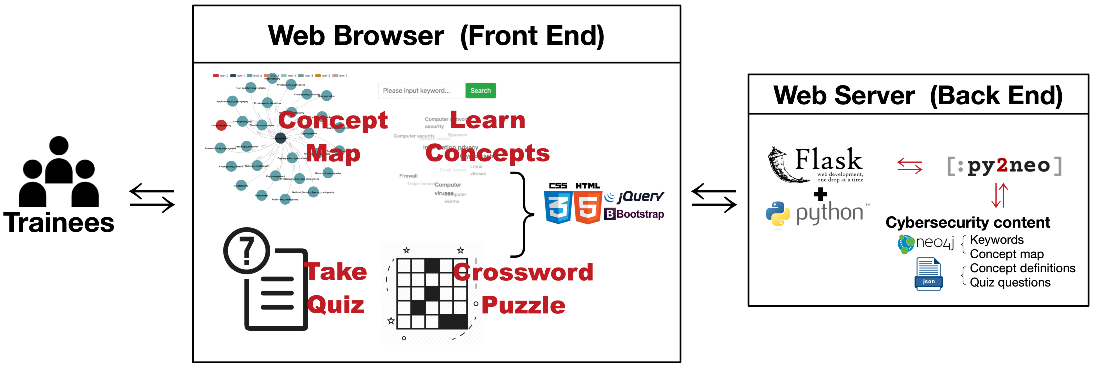


Figure 4.1: The architecture of CyATP.

An overview of the CyATP framework is provided in the figure 4.1. Trainees use the web interface to access the Concept Map and Learn Concepts pages to find out about the security concepts they want to study. They can also use a Take Quiz and a Crossword Puzzle to test and deepen their knowledge. The front end of the CyATP platform is developed using Bootstrap and jQuery, and the back end employs Flask and Neo4j database.

This platform is roughly divided into two parts: the learning activity component and the serious game component.

4.2 Learning Activity Component

In this section will focus on the concept map and learning page in the learning activity component, understand what functions they have and how to provide the trainees with cybersecurity related learning content.

4.2.1 Concept Map

The screenshot of the concept map page is shown in the figure 4.2.

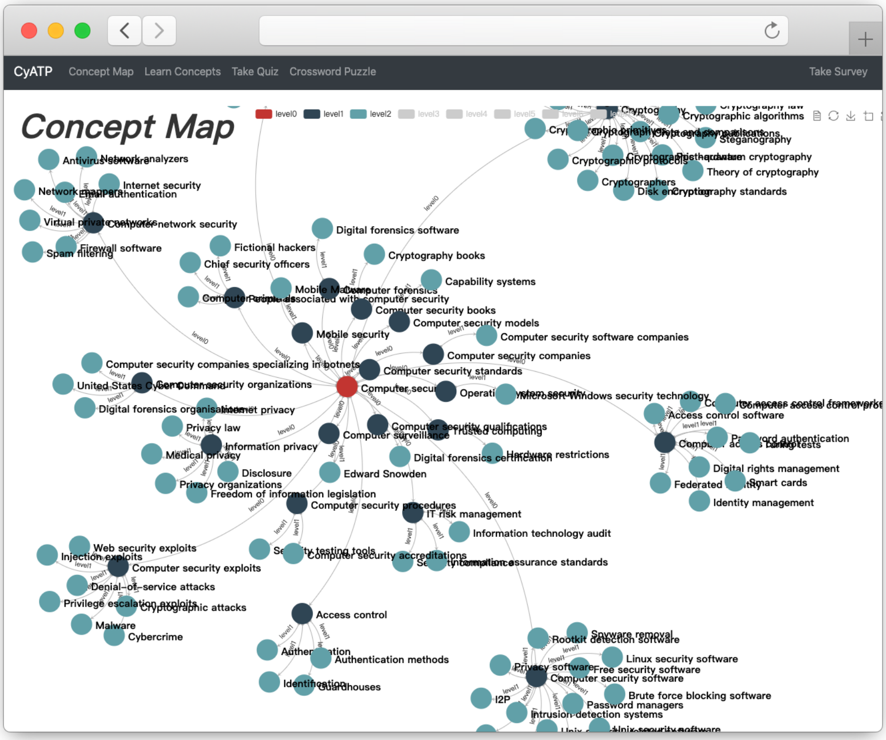


Figure 4.2: The screenshot of the concept map page.

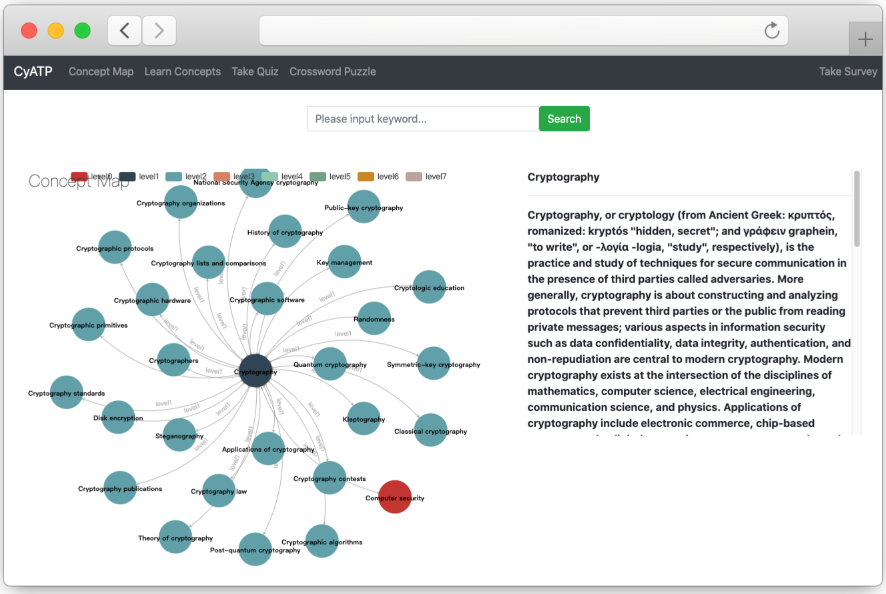
Trainees can understand the overall structure of cybersecurity knowledge through this diagram. The hierarchical structure of the concept map presents numerous concepts and their relationships in the form of visualization. The relationship between keywords and keywords are highly condensed, making the links between knowledge no longer complicated. Trainees can easier to master and remember the knowledge.

The nodes of each layer are represented by different colors. Click the level button in the upper navigation bar to display or undisplay the corresponding layer's nodes. When the mouse is pointing on a node, it will focus on the node that the node points to or is pointed to. Use interaction to learn the knowledge, stimulate learners' interest, and enable them to discover and explore more knowledge continuously.

Here used security content from collecting dataset in the section 3.2.2.

4.2.2 Learn Concepts Page

The screenshot of the learn concepts page is shown in the figure 4.3.



Cryptography

Cryptography, or cryptology (from Ancient Greek: κρυπτός, romanized: kryptós "hidden, secret"; and γράφειν graphein, "to write", or -λογία -logia, "study", respectively), is the practice and study of techniques for secure communication in the presence of third parties called adversaries. More generally, cryptography is about constructing and analyzing protocols that prevent third parties or the public from reading private messages; various aspects in information security such as data confidentiality, data integrity, authentication, and non-repudiation are central to modern cryptography. Modern cryptography exists at the intersection of the disciplines of mathematics, computer science, electrical engineering, communication science, and physics. Applications of cryptography include electronic commerce, chip-based

Figure 4.3: The screenshot of the learn concepts page.

Trainers can search for cybersecurity related keywords they are interested in through this page, or select a topic of interest through the word cloud under the search bar. After that, the concept map related to the word will be displayed on the page's left side. By clicking the node in the concept map, the corresponding concept text will be displayed on the page's right side. After clicking, the concept text after clicking will be saved below the display box on the right to facilitate users to browse the previous records.

This page also provides the fuzzy search. For example, when the trainee enters "security" in the search bar, the page will recommend knowledge related to "security", such as "Data security", "Mobile security", "Internet security", and "Database security".

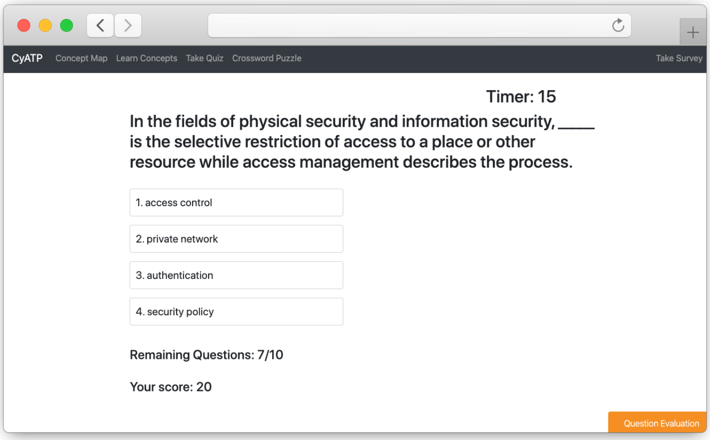
This exploratory learning based on interest can reduce the boredom of learning, make trainees spend their energy on the knowledge they don't know, and promote knowledge retention.

4.3 Serious Games Component

In this section we will discuss the quiz and crossword puzzle game in the serious game component, to know about the game elements and mechanisms, and the corresponding cybersecurity learning content provided in the game.

4.3.1 Quiz Game

The figure 4.4 shows the screenshot of the quiz game.



Timer: 15

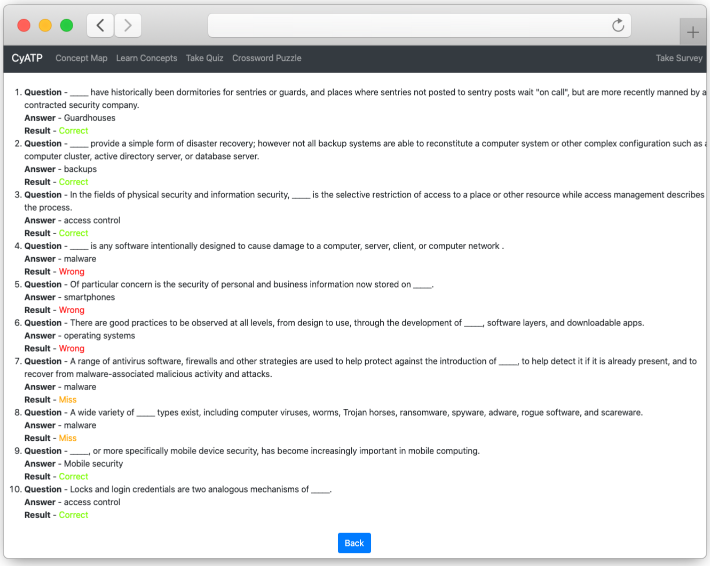
In the fields of physical security and information security, ____ is the selective restriction of access to a place or other resource while access management describes the process.

1. access control
2. private network
3. authentication
4. security policy

Remaining Questions: 7/10

Your score: 20

Figure 4.4: The screenshot of the quiz page.



1. **Question** - ____ have historically been dormitories for sentries or guards, and places where sentries not posted to sentry posts wait "on call", but are more recently manned by a contracted security company.
   Answer - Guardhouses
   Result - Correct
2. **Question** - ____ provide a simple form of disaster recovery; however not all backup systems are able to reconstitute a computer system or other complex configuration such as computer cluster, active directory server, or database server.
   Answer - backups
   Result - Correct
3. **Question** - In the fields of physical security and information security, ____ is the selective restriction of access to a place or other resource while access management describes the process.
   Answer - access control
   Result - Correct
4. **Question** - ____ is any software intentionally designed to cause damage to a computer, server, client, or computer network .
   Answer - malware
   Result - Wrong
5. **Question** - Of particular concern is the security of personal and business information now stored on ____.
   Answer - smartphones
   Result - Wrong
6. **Question** - There are good practices to be observed at all levels, from design to use, through the development of ____, software layers, and downloadable apps.
   Answer - operating systems
   Result - Wrong
7. **Question** - A range of antivirus software, firewalls and other strategies are used to help protect against the introduction of ____, to help detect it if it is already present, and to recover from malware-associated malicious activity and attacks.
   Answer - malware
   Result - Miss
8. **Question** - A wide variety of ____ types exist, including computer viruses, worms, Trojan horses, ransomware, spyware, adware, rogue software, and scareware.
   Answer - malware
   Result - Miss
9. **Question** - ____, or more specifically mobile device security, has become increasingly important in mobile computing.
   Answer - Mobile security
   Result - Correct
10. **Question** - Locks and login credentials are two analogous mechanisms of ____.
    Answer - access control
    Result - Correct

Figure 4.5: The screenshot of the feedback of quiz page.

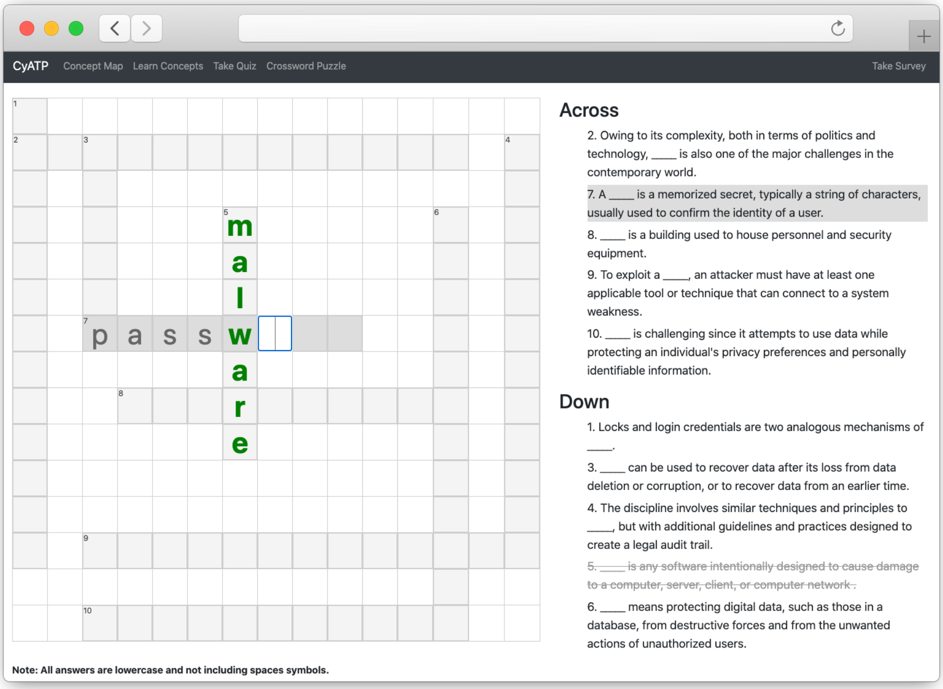
There are 10 questions in each round, and they are randomly selected from the generated content. There is a 30-second answer event for each question. For each correct answer, 10 points are awarded, and no points are deducted for incorrect answers. The remaining questions and scores will be displayed below each question. Every time trainees complete a question, they can immediately know whether the answer is correct or incorrect, and at the same time give the correct answer.

After all the questions are answered, the overall score will be reported to the trainees. The trainees can view all the questions in the current round and their own answers by clicking the "more details" button, can be see in figure 4.5 on the previous page. There is a "Question Evaluation" button below all questions. Since all questions are generated by Machine Learning, if there are questions with different semantics, the trainees can use this button to feedback on the current question. Developers make improvements through this users' feedback.

The quiz method can combine the cybersecurity learning content with the game well, and attract trainees to participate in the training by scoring. The way of feedback at any time enables trainees to know their own shortcomings, have more confidence in what they know, and deepen their understanding of the correct content. The time pressure allows the trainees to consider in a limited time, preventing delay and cheating.

4.3.2 Crossword Puzzle Game

The figure 4.6 shows the screenshot of the crossword puzzle game.



Across

2. Owing to its complexity, both in terms of politics and technology, ____ is also one of the major challenges in the contemporary world.

7. A ____ is a memorized secret, typically a string of characters, usually used to confirm the identity of a user.

8. ____ is a building used to house personnel and security equipment.

9. To exploit a ____, an attacker must have at least one applicable tool or technique that can connect to a system weakness.

10. ____ is challenging since it attempts to use data while protecting an individual's privacy preferences and personally identifiable information.

Down

1. Locks and login credentials are two analogous mechanisms of ____.

3. ____ can be used to recover data after its loss from data deletion or corruption, or to recover data from an earlier time.

4. The discipline involves similar techniques and principles to ____, but with additional guidelines and practices designed to create a legal audit trail.

~~5. ____ is any software intentionally designed to cause damage to a computer, server, client, or computer network .~~

6. ____ means protecting digital data, such as those in a database, from destructive forces and from the unwanted actions of unauthorized users.

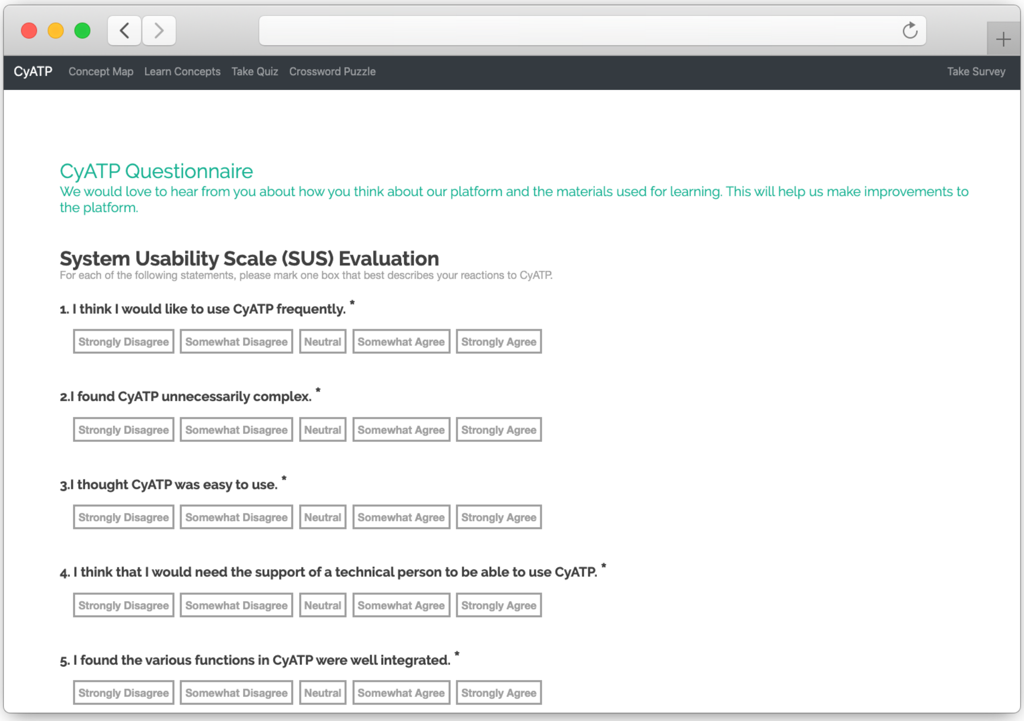
Note: All answers are lowercase and not including spaces symbols.

Figure 4.6: The screenshot of the crossword puzzle page.

On the left is the crossword puzzle part, and on the right is the clues. The game's rule is to solve the puzzle based on limited clues, then calculate the next answer based on the guessed answer and the hints given by the maker until space is completely filled. The clues come from short questions generated by the content, and the answers come from keywords. This game needs all input in lowercase and do not include space symbols. When a puzzle is solved, the clue part on the right will turn into a gray underline and the input answer will turn into green. We provided a total of 10 puzzles, two of which are 13×13, one of which is 13×14, three of which are 14×14, three of which are 15×15, and one of which is 16×16. Developers can also generate more different forms of puzzles based on the source code we provide.

Playing the crossword puzzle is the process of solving the problem. Players will get some clues, which are difficult to understand and easily cause confusion. The brain must jump out of the original thinking framework and think in different ways to solve the puzzle. In the process of analyzing clues, can exercise reasoning ability and cultivate concentration. While playing this game, that can also learn cybersecurity content and balance learning and entertainment.

In this web application, trainees can not only feedback the quality of the generated content, but also evaluate our platform. As shown in the figure 4.7, we have specially designed an evaluation page, hoping that trainees will participate in this survey so that we can understand their needs and the inadequacies of the platform, help us improve the platform, and provide them with a better user experience.



CyATP Questionnaire

We would love to hear from you about how you think about our platform and the materials used for learning. This will help us make improvements to the platform.

System Usability Scale (SUS) Evaluation

For each of the following statements, please mark one box that best describes your reactions to CyATP.

1. I think I would like to use CyATP frequently. *

Strongly Disagree | Somewhat Disagree | Neutral | Somewhat Agree | Strongly Agree

2.I found CyATP unnecessarily complex. *

Strongly Disagree | Somewhat Disagree | Neutral | Somewhat Agree | Strongly Agree

3.I thought CyATP was easy to use. *

Strongly Disagree | Somewhat Disagree | Neutral | Somewhat Agree | Strongly Agree

4. I think that I would need the support of a technical person to be able to use CyATP. *

Strongly Disagree | Somewhat Disagree | Neutral | Somewhat Agree | Strongly Agree

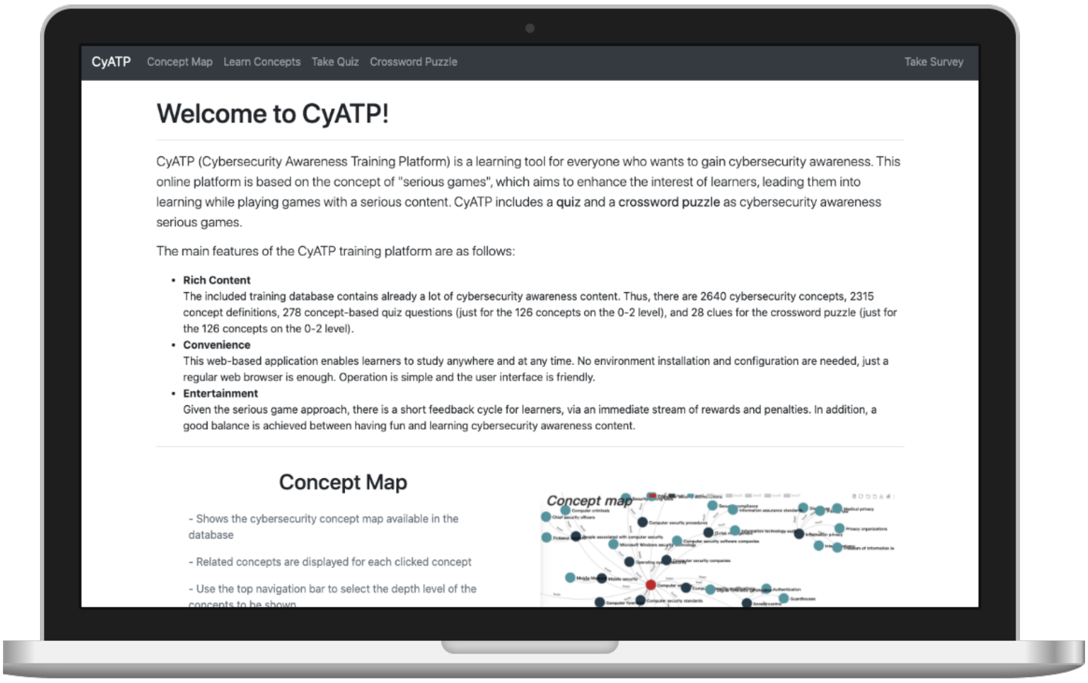
5. I found the various functions in CyATP were well integrated. *

Strongly Disagree | Somewhat Disagree | Neutral | Somewhat Agree | Strongly Agree

Figure 4.7: The screenshot of the survey page.

4.4 Implementation

Our implementation of this online web application awareness training platform is mainly divided into two parts: the front end and the back end. The user interacts with the front end web page; then, the back end receives the request and retrieves the corresponding data from the database, returns it to the front end; finally, the front end displays the corresponding content to the user.



Welcome to CyATP!

CyATP (Cybersecurity Awareness Training Platform) is a learning tool for everyone who wants to gain cybersecurity awareness. This online platform is based on the concept of "serious games", which aims to enhance the interest of learners, leading them into learning while playing games with a serious content. CyATP includes a **quiz** and a **crossword puzzle** as cybersecurity awareness serious games.

The main features of the CyATP training platform are as follows:

- **Rich Content**
  The included training database contains already a lot of cybersecurity awareness content. Thus, there are 2640 cybersecurity concepts, 2315 concept definitions, 278 concept-based quiz questions (just for the 126 concepts on the 0-2 level), and 28 clues for the crossword puzzle (just for the 126 concepts on the 0-2 level).
- **Convenience**
  This web-based application enables learners to study anywhere and at any time. No environment installation and configuration are needed, just a regular web browser is enough. Operation is simple and the user interface is friendly.
- **Entertainment**
  Given the serious game approach, there is a short feedback cycle for learners, via an immediate stream of rewards and penalties. In addition, a good balance is achieved between having fun and learning cybersecurity awareness content.

Concept Map

- Shows the cybersecurity concept map available in the database

- Related concepts are displayed for each clicked concept

- Use the top navigation bar to select the depth level of the concepts to be shown

Figure 4.8: The screenshot of the CyATP home page.

- Front end part: The CyATP is developed web application. Mainly use the open source framework Bootstrap and jQuery to design the front-end page. The application can not only be used on the computer side, but also adapted to the screen resolution of the mobile phone.
- Back end part: CyATP uses the lightweight python web framework Flask. Flask is based on the Werkzeug WSGI toolkit and Jinja2 template engine. We store keywords and concept maps in the relational database Neo4j, and store the generated questions and puzzle data in a JSON file. Use Neo4j from within a Python application and from the command line via the python library py2neo.

Chapter 5

Evaluation

This chapter will evaluate from two aspects: training content and training platform. First discuss the evaluation methods including participants and procedure, and then discuss and analyze the evaluation results.

5.1 Method

Participants

A questionnaire survey was conducted with ten master students from the Department of Information Science, Knowledge Science and Martial Science, Japan Advanced Institute of Science and Technology school. With the gender split of 50% male and 50% female. 70% of the participants have experience in security awareness training, but 80% of them self-assessed their information security knowledge level as beginners. All participants who take part in the survey are voluntary and unpaid. A summary of the demographics of the participants is shown in table 5.1 on the following page.

Procedure

The questionnaire is conducted by each person one by one.

First, explain the experiment's purpose and process to the trainee. They were also informed that they can make any comment or feedback during the experiment, or even withdraw from the experiment without giving any reason for withdrawing.

Then ask the trainee to read the experiment introduction and fill in personal information anonymously. One item is to ask the trainee whether he/she has participated in similar cybersecurity awareness training in the personal information form. If they gave an affirmative answer, ask them what kind of cybersecurity awareness training they were participating in and gave them a brief introduction of what cybersecurity awareness training is to confirm they understand, while negative responders were read the definition of cybersecurity awareness training and gave them a short description. After that, let them evaluate their level of cybersecurity knowledge.

Characteristics	**Total**
Number of people	10
Gender	
Male	5
Female	5
Degree	
Master student	10
Major	
Information science	5
Knowledge science	3
Material science	2
Security awareness training experience	
experienced	7
inexperienced	3
Self-assessment about cybersecurity knowledge	
Beginner	8
Intermediate	2

Table 5.1: The demographics of participant.

Next, we give trainees about 20 minutes to use our platform for training. After the training is completed, take about 20 minutes to fill in three questionnaires (for details about the questionnaire can refer to the appendix), which are evaluation of platform, evaluation of serious games and evaluation of learning material quality.

After finishing all questionnaires, trainees were thanked for their support and valuable time participating in this research.

5.2 Training Content Evaluation

For the questionnaire to evaluate the training content, can refer to Questionnaire #3 in the appendix. There are 8 questions in the questionnaire. Participants need to rate the scale from 1 (strongly disagree) to 5 (strongly agree).

The questions are roughly divided into 4 groups. The first group(Q1-Q4) is about evaluating the learning material in the learning activity component. Use these questions to investigate whether the knowledge covered by the learning materials provided can satisfy the learner, whether it is easy to understand, and whether the concept map can help to learn. The second group(Q5-Q7) is about evaluating the generated content in the serious game component. Confirm whether the generated question has grammatical errors, whether the choices are appropriate, and whether the clues provided reflect the answer. The third group (Q8) wanted to know how satisfied they were with the learning materials provided in general. The last question is an open question to asked participants to write their any comments or suggestions.

Evaluation results of training content are shown in table 5.2.

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Score**	**Score Average**
Trainee 1	5	5	5	5	3	4	5	5	37	34.4
Trainee 2	5	4	4	4	4	5	3	4	33	
Trainee 3	5	5	5	4	4	4	3	3	33	
Trainee 4	5	5	5	5	5	5	5	5	40	
Trainee 5	5	5	5	4	4	4	3	5	35	
Trainee 6	5	5	5	5	5	5	5	5	40	
Trainee 7	4	5	4	3	3	4	4	4	31	
Trainee 8	4	4	4	4	4	4	2	4	30	
Trainee 9	5	5	4	3	5	5	5	4	36	
Trainee 10	5	4	3	3	3	4	3	4	29	
Question Average	4.8	4.7	4.4	4	4	4.4	3.8	4.3		
Question Variance	0.16	0.21	0.44	0.6	0.6	0.34	1.16	0.41		

Table 5.2: The evaluation results of training content.

All the trainees gave an average score of 34.4 points, of which the highest score was 40 points, and the lowest score was 29 points. All scores given are higher than 24 points (select neutral for each question).

The trainees' evaluation of the concept map's learning content is very high (the average value of Q1 is 4.8, the average value of Q2 is 4.7). And think that the concept text is easy to understand and suitable for learning (Q3 average is 4.4, Q4 average is 4). The average score of those three questions (Q5-Q7) is 4.067, which is higher than 3 (select neutral for each question), that evaluation of the content generated by NLG. It indicates that the generated content is clear and roughly grammatically correct. But some trainees gave high evaluations about the quality of the relevant learning content in serious games, but some remained neutral. The reason is that some trainees are not native English speakers. When playing quiz, they don't pay too much attention to the question's grammatical problems and cannot give a positive evaluation.

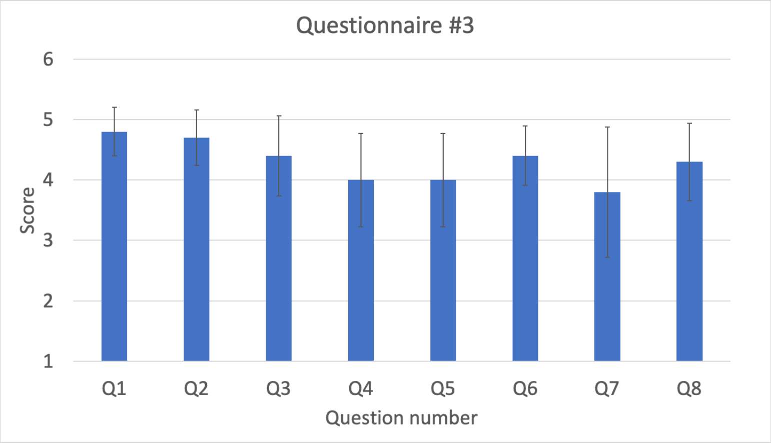


Figure 5.1: Barplot of the evaluation results of training content.

As shown in the figure 5.1, the variance line of Q7 is relatively long. For trainers who are not native English speakers, the crossword puzzle may be difficult for them due to insufficient vocabulary. Combining professional knowledge with this game increases the game's challenge even more, making it difficult for some trainees to analyze the answers given clues.

Generally speaking, the trainees are satisfied with the training content provided and think that it can help them learn cybersecurity knowledge and improve security awareness.

5.3 Training Platform Evaluation

5.3.1 Platform Evaluation

Performance Evaluation

We use a web page front-end performance testing tool WebPageTest [32] to test the performance of our platform. This online free performance evaluation website that supports IE, Chrome and other real browsers, and real consumer connection speeds run free website speed tests from multiple locations worldwide. That can choose to run the simple tests or perform advanced tests, including multi-step transactions, video capture, content blocking, and more. A wealth of diagnostic information will also be provided based on the test results, and a final rating will be given to each item.

We used Advanced Testing, where the test location in Tokyo, Japan, and Chrome was used for 3 rounds of testing. Test items include:

- First Byte Time: refers to the time when the browser receives the first byte of HTML content, including DNS lookup, TCP connection, SSL negotiation (if it is an HTTPS request) and TTFB (Time To First Byte).
- Keep-alive Enabled: Requesting the content on the webpage needs to establish a connection with the web server. It will take much time to reconnect with each

request. If the keep-alive method is used, the time to load the page can be reduced by 40%-50%.

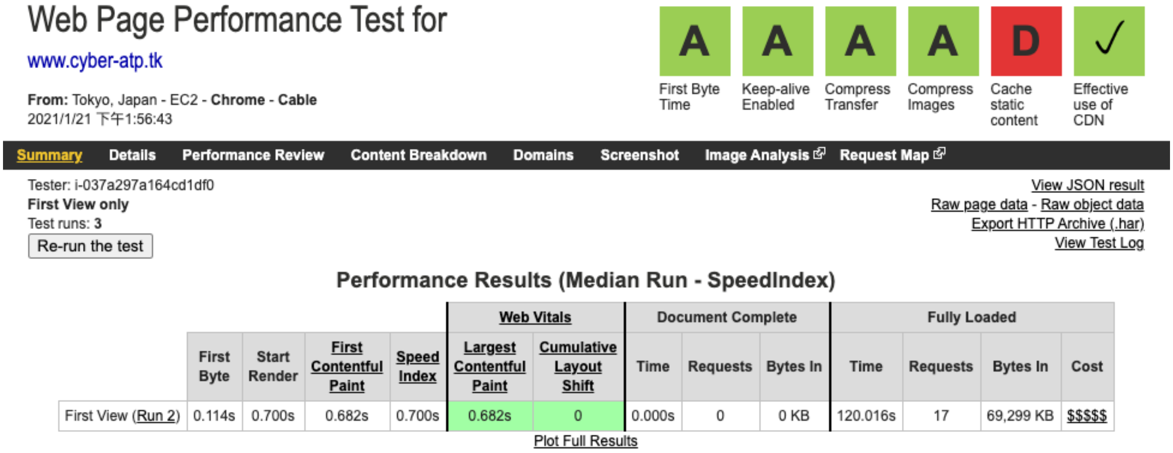
- Compress Transfer: Because text resources are usually loaded at the beginning of the page, if using the HTTP compression method of transmission can improve performance and reduce the amount of data transmitted.

- Compress Images: Make sure that the photo's quality is not set too high, which will cause the resource to load very slowly. Usually, use JPEG images, which can be compressed to a large extent while ensuring the visual quality.

- Cache Static Content: browsers store static content that is not frequently changed in the cache, reducing the burden on the web server.

- Effective Use of CDN: The content distribution network can redirect the user's request to the service node closest to the user in real time based on comprehensive information such as network traffic and the connection of each node, load status, distance to the user and response time. Its purpose is to enable users to obtain the required content nearby, solve the congestion of the Internet network, and improve users' response speed visiting the website.

The test results are shown in the figure 5.2. We reached the highest rating of A in the four projects (First byte, Keep-alive enabled, Compress transfer, Compress images), and used the CDN efficiently, but the cache static content score was not high as D (66%). The reason is that we use GIF format animated pictures in the web pages, which are large in size and inconvenient to compress, which occupies the web cache and makes the rating lower. We will adjust the structure of the application in future work to improve the user experience.



Web Page Performance Test for

www.cyber-atp.tk

From: Tokyo, Japan - EC2 - **Chrome** - **Cable**
2021/1/21 下午1:56:43

A	A	A	A	D	✓
First Byte Time	Keep-alive Enabled	Compress Transfer	Compress Images	Cache static content	Effective use of CDN

Tester: i-037a297a164cd1df0
First View only
Test runs: **3**

Performance Results (Median Run - SpeedIndex)

					Web Vitals		Document Complete			Fully Loaded			
	First Byte	Start Render	First Contentful Paint	Speed Index	Largest Contentful Paint	Cumulative Layout Shift	Time	Requests	Bytes In	Time	Requests	Bytes In	Cost
First View (Run 2)	0.114s	0.700s	0.682s	0.700s	0.682s	0	0.000s	0	0 KB	120.016s	17	69,299 KB	$$$$$

Figure 5.2: The evaluation results of web page test.

User Evaluation

We used the SUS (System Usability Scale) [33] to evaluate the usability of the CyATP platform. For specific questions, can refer to questionnaire #1 in the appendix. This Scale provides a "quick and dirty", reliable tool for measuring usability. There are 10 questions in total, and the trainees will rate them according to their level of agreement with the statements they read. Each question will be graded from 1 (strongly disagree) to 5 (strongly agree). Every odd-numbered question is a positive question, and every even-numbered question is a negative question. The questions in the questionnaire mainly measure three aspects: effectiveness, efficiency and satisfaction.

The SUS is easy to manage and is a short evaluation test. There are free templates on the Internet, so there is no need to re-research and design. This scale has been used for nearly 30 years and is used to evaluate various products and services, including hardware, software, applications, etc., and has a certain degree of universality and authority.

The SUS points rules are as follows:

- For each odd-numbered question, subtract 1 from the score given by the trainer.

$$O = \sum_{i=1,3,5,7,9} (Score_i - 1)$$

- For each even-numbered question, subtract the score given by the trainer from 5.

$$E = \sum_{i=2,4,6,8,10} (5 - Score_i)$$

- Add the above scores and multiply by 2.5.

$$FinalScore = (O + E) * 2.5$$

The full score is 100 points, and this article [34] points out that the average SUS score based on 500 studies is 68 points.

The evaluation results of CyATP according to SUS are shown in the table 5.3 on the following page. The average score is 80.5 points, the highest score is 100 points, and the lowest score is 45 points. Two trainees evaluated below 60 points. They commented that they hope to produce a multi-language version of the website to facilitate their learning of cybersecurity knowledge. Only the English version is complicated for non-native speakers to learn professional knowledge. We will improve this in future work.

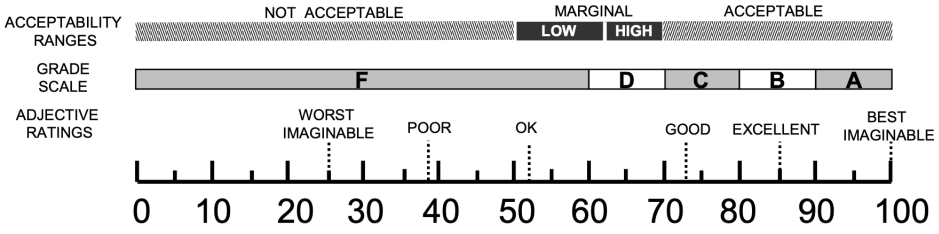


Figure 5.3: The score of SUS [35].

No	**Trainee 1**	**Trainee 2**	**Trainee 3**	**Trainee 4**	**Trainee 5**	**Trainee 6**	**Trainee 7**	**Trainee 8**	**Trainee 9**	**Trainee 10**
Q1	4	5	3	5	5	4	4	2	5	3
Q2	2	1	2	1	1	1	2	3	2	2
Q3	4	5	4	5	5	5	5	3	5	4
Q4	1	2	4	1	1	1	1	4	3	2
Q5	4	5	4	5	5	5	3	2	5	3
Q6	4	1	2	1	3	1	1	4	2	1
Q7	5	4	5	5	4	5	4	4	5	4
Q8	2	1	2	1	1	1	1	3	1	1
Q9	5	5	2	5	4	5	4	5	4	4
Q10	1	2	5	1	1	1	2	4	2	2
Score	80	92.5	57.5	100	90	97.5	82.5	45	85	75
Average Score	80.5									

Table 5.3: The evaluation results of platform.

As the figure 5.3 on the previous page shows the SUS score, that including the adjective ratings (worst imaginable, poor, ok, good, excellent, best imaginable), acceptability scores (not acceptable, marginal, acceptable), and school grading scales (F, D, C, B, A).

According to the average score of 80.5 given by the trainees and the standard form, CyATP is in the B grade, which is a good and acceptable platform for cybersecurity awareness training. We will also continue to improve and develop based on the comments given by the trainees.

5.3.2 Serious Games Evaluation

The update of the game is very rapid, and the development of serious games also changes with the game changes. The evaluation of serious games does not have a clear industry standard like software engineering. The entertainment part and the serious part are not only the difficulty of developing serious games, but also the difficulty of evaluating serious games.

We used an evaluation scale tool developed by Emmanual Fokides [36] to check the effectiveness of serious games while contrasting the user's perspective. The evaluation scale was obtained by 542 students playing two serious games and analyzing the questionnaires they filled out through exploratory and confirmatory factor analysis. The final version of the serious game evaluation scale uses 12 factors and 53 items to evaluate serious games' satisfactory reliability and validity.

Since this serious game evaluation scale is developed for all serious games, some factors are not suitable for evaluating CyATP. For example, realism is used to evaluate the immersion of 3D games, auditory adequacy is used to evaluate the sound effects and images of video games, and narration is used to evaluate the game's story. We ultimately retained 9 factors and 29 items to evaluate our serious games in CyATP. For specific questions, please refer to Questionnaire #2 in the appendix.

The 9 factors are specifically: presence (Q1-Q4), enjoyment (Q5-Q8), learning effectiveness (Q9-Q14), feedback (Q15), relevance to (Q16-Q17), goal's clarity (Q18), ease of use (Q19-Q23), adequacy of the learning material (Q24-Q26), motivation (Q27-Q29).

There is a total of 29 questions, of which 6, 20, 21, 22, 24, 25, 27, 28, 29 are negative questions, and the rest are positive questions. Trainees need to rate them according to their level of agreement with the statements they read, from 1 to 5, with 1 being strongly disagree and 5 being strongly agree.

The results of the questionnaire survey of serious games (quiz and crossword puzzle) evaluation are shown in the table 5.4 on the following page.

No	Trainee 1	Trainee 2	Trainee 3	Trainee 4	Trainee 5	Trainee 6	Trainee 7	Trainee 8	Trainee 9	Trainee 10	Average	Factor Average	Factor Variance
Q1	5	4	2	5	4	5	4	5	5	4	4.3	3.9	0.115
Q2	4	4	3	5	4	2	3	3	3	3	3.4		
Q3	4	5	2	5	4	5	4	4	4	4	4.1		
Q4	5	3	2	5	4	5	4	3	4	3	3.8		
Q5	4	4	3	5	3	4	4	4	3	3	3.7	4.05	0.1125
Q6	4	4	3	5	3	5	4	4	3	4	3.9		
Q7	5	4	2	5	3	5	4	3	5	4	4		
Q8	5	5	4	5	4	5	5	3	5	5	4.6		
Q9	5	4	4	5	3	5	5	4	4	4	4.3	4.3333	0.0722
Q10	4	5	4	5	3	5	4	4	3	4	4.1		
Q11	4	4	4	5	3	5	4	4	4	4	4.1		
Q12	5	5	5	5	5	5	5	4	3	5	4.7		
Q13	5	5	5	5	5	5	5	4	4	4	4.7		
Q14	5	4	3	5	5	5	4	4	3	3	4.1		
Q15	4	5	5	5	5	5	5	4	5	5	4.8	4.8	0
Q16	5	3	4	5	4	5	4	4	5	3	4.2	4.05	0.0225
Q17	5	3	4	5	3	5	4	4	3	3	3.9		
Q18	5	5	4	5	5	5	4	3	4	4	4.4	4.4	0
Q19	5	5	3	5	5	4	4	5	2	5	4.3	4.18	0.0376
Q20	4	5	4	5	5	5	5	1	3	5	4.2		
Q21	5	5	4	5	2	5	4	2	2	4	3.8		
Q22	5	4	3	5	5	5	5	2	5	4	4.3		
Q23	4	4	4	5	4	5	4	4	5	4	4.3		
Q24	2	3	2	5	3	5	3	1	3	2	2.9	3.6333	0.3289
Q25	4	4	3	5	3	5	4	2	5	2	3.7		
Q26	5	4	5	5	4	5	4	4	3	4	4.3		
Q27	2	5	3	5	3	5	5	2	5	4	3.9	3.9333	0.0422
Q28	5	4	3	5	3	5	4	3	1	4	3.7		
Q29	5	1	5	5	4	5	5	3	4	5	4.2		

Table 5.4: The evaluation results of serious games.

Among all the question items, the highest average score of Q15 is 4.8. Most trainees think that our serious games have a very fast response speed and immediate feedback on their operations. Tied for second place are Q12 and Q13, with an average score of 4.7. Through our two serious games, the trainees think that they have learned some basic cybersecurity knowledge in the game and increased their cybersecurity awareness. Q8 ranked third, with an average score of 4.6. Whenever trainees complete each round of the quiz, they feel good about what they have done.

As the figure 5.4 shows the evaluation results of the serious games on 9 factors.

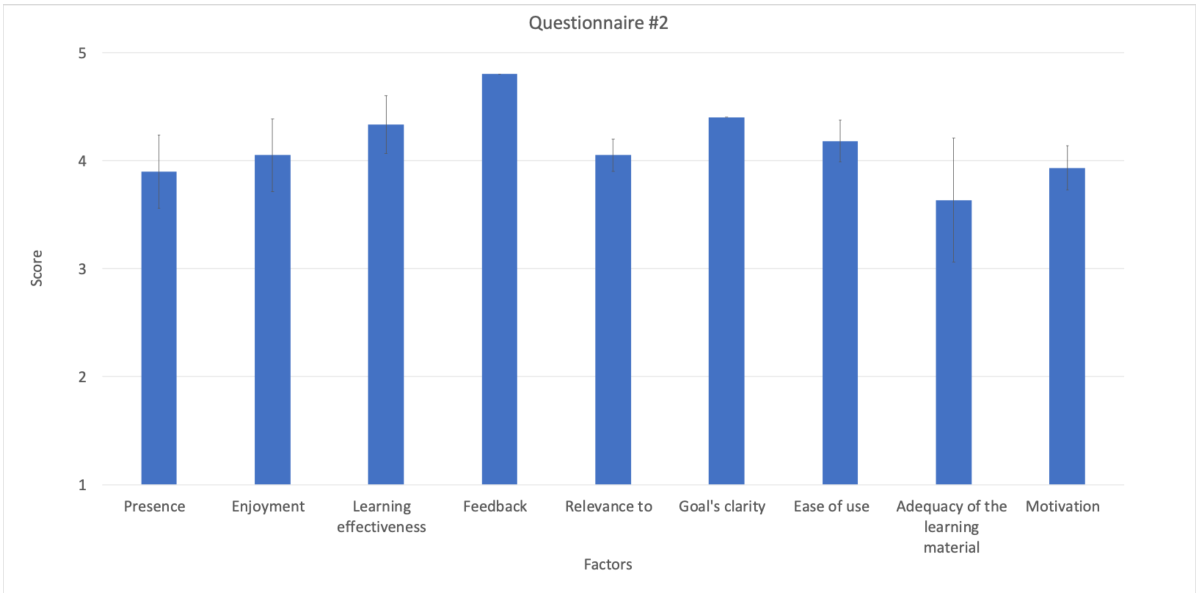


Figure 5.4: The evaluation results of serious games.

Among the 9 factors, feedback is the most prominent, and the average score of this factor is 4.8. The second is clear goals with an average score of 4.4 and the third is learning effectiveness, with an average score of 4.33. Among them, the lowest average score of the learning materials factor's adequacy is 3.63, and the highest variance value is 0.329. However, it also exceeds the 3 points of neutral choice for all questions. The item in this factor Q24 score is the lowest, many trainees think that in some cases too much information is given and it is not easy to remember essential knowledge. The reason is that on the concept map page, we display the trainees all the concepts related to cybersecurity from the beginning, which caused the trainees to receive a lot of information at once and felt pressure. In future development work, we will adjust the amount of information displayed to attract trainees first and arouse their interest instead of stressing them.

The trainees' personal experience of our serious games and the evaluation shows that those serious games are easy to use, give users immediate feedback, have clear goals, and efficiently learn security knowledge while playing the game.

Chapter 6

Conclusion and Future Work

6.1 Concluding Remarks

In the first half of 2020, a global outbreak of COVID-19 began. To prevent the spread of the virus, people began to work at home and reduce travel. It was at this time that I started this research. When schools and companies began to implement remote workforce measures, many news reports increased many cybersecurity risks to the organization. Cybercriminals use phishing emails related to COVID-19 to flood employees' inboxes, and seemingly harmless attachments are malicious software that lures unsuspecting employees to open. In such a social environment, cybersecurity awareness training becomes more and more necessary and important.

Since the COVID-19, sports enthusiasts have been restricted from going out for sports. However, a Ring Fit Adventure game developed by Nintendo is popular with many people. Using the combination of serious games and fitness, players can enjoy the entertainment brought by the game, and they can also exercise. Not only serious games are used in health, but also cyber security awareness training.

Based on the current social environment, we have developed a web application including exploratory interest learning and serious games, hoping to provide a useful learning and training platform. At the same time, whether it is education or training, it is always unavoidable to solve the problem of the source of learning content. We use natural language generation to automatically generate learning content, reduce development costs, and generate large amounts of educational materials efficiently and quickly.

Finally, we invited some volunteers to use our platform and they gave us feedback afterwards. The trainees' evaluation of the concept map's learning content is a very high score, and they think that the concept text is easy to understand and suitable for learning. We used the SUS (System Usability Scale) to evaluate the usability of the CyATP platform. According to the average score of 80.5 given by the trainees, CyATP is in the B grade, a good and acceptable platform for cybersecurity awareness training.

About the evaluation of serious games, we used 9 factors and 29 items. The trainees' personal experience of our serious games and the evaluation shows that those serious games are easy to use, give users immediate feedback, have clear goals, and efficiently learn security knowledge while playing the game. Generally speaking, our CyATP application can help users learn relevant security knowledge and improve their awareness, and put it to use in their daily life.

6.2 Future Work

Although our platform has many strengths, it is not perfect. The following points may be improved in the future:

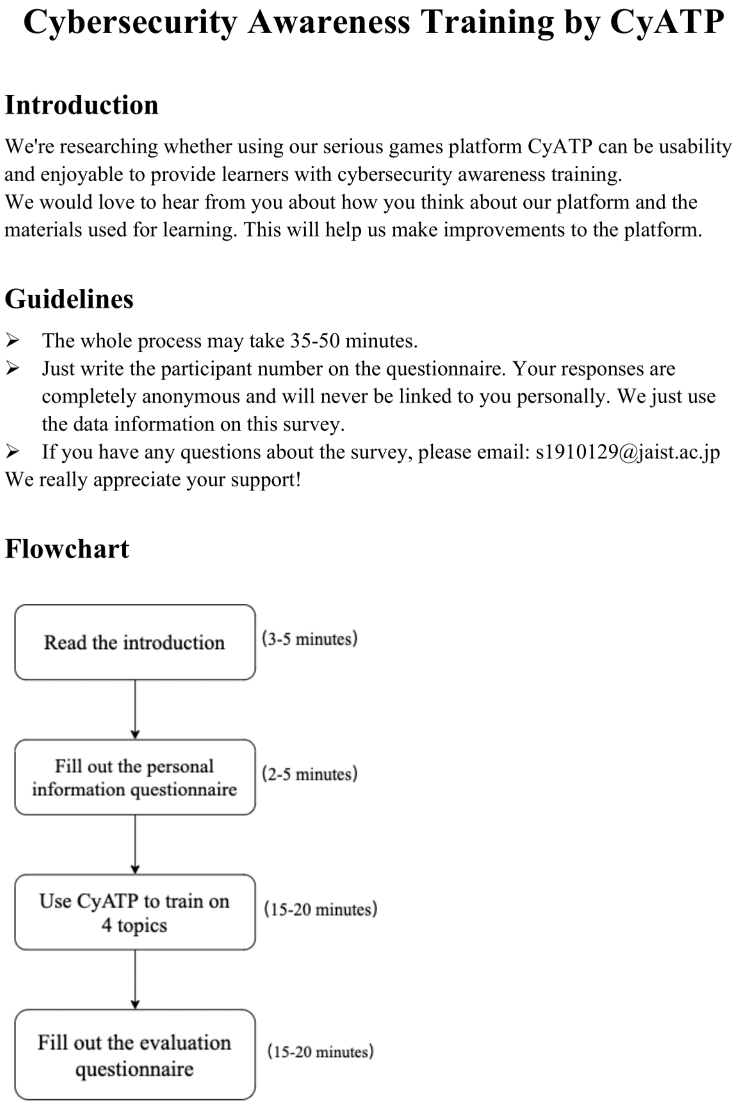
- *Combine CyATP with an adaptive recommendation system.* Now CyATP provides learning and serious gaming services to all visitors, without recording any user information or record. In the subsequent development, we hope to add a login entry to provide personalized services for trainees. Based on the trainee's historical behavior, make personalized recommendations to them. For example, users spend more time in serious games than exploratory interest learning. The system uses the serious game method to provide more content to attract them to continue training. Taking into account the timeliness of historical behavior, the system also provides the function of reviewing learning materials. Based on the content that the trainee has learned, recommendations are related to these contents, satisfying their curiosity while understanding their needs.

- *Optimize CyATP's web application structure.* Due to limited development time and unfamiliarity with application development, our platform still needs to be adjusted in architecture. As mentioned in Section 5.3.1, our platform did not perform well on cache static content. We will use lossy compress GIF animations to reduce cache pressure. In the future, we will unify all data structures and store all data in the database to ensure data security and efficient access. When users reach a certain scale, consider adding high-concurrency modules, use caching reasonably, optimize database query and access, and provide users with adequate services.

- *Support training content generation in other languages.* At present, our platform only provides English learning content and environment, which may increase the difficulty for non-native speakers to learn related cybersecurity professional knowledge. To assist more users who want to raise security awareness through our platform, we plan to develop security learning content in multiple languages to lower the learning threshold, and make it easier and more convenient for more people to use.

Appendix A

Questionnaire



Cybersecurity Awareness Training by CyATP

Introduction

We're researching whether using our serious games platform CyATP can be usability and enjoyable to provide learners with cybersecurity awareness training.
We would love to hear from you about how you think about our platform and the materials used for learning. This will help us make improvements to the platform.

Guidelines

- The whole process may take 35-50 minutes.
- Just write the participant number on the questionnaire. Your responses are completely anonymous and will never be linked to you personally. We just use the data information on this survey.
- If you have any questions about the survey, please email: s1910129@jaist.ac.jp

We really appreciate your support!

Flowchart

Read the introduction (3-5 minutes)

Fill out the personal information questionnaire (2-5 minutes)

Use CyATP to train on 4 topics (15-20 minutes)

Fill out the evaluation questionnaire (15-20 minutes)

Questionnaire #1

Cybersecurity Awareness Training by CyATP

System Usability Scale (SUS) Evaluation

For each of the following statements, please mark one box that best describes your reactions to CyATP today.

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
1. I think I would like to use CyATP frequently.	□	□	□	□	□
2. I found CyATP unnecessarily complex.	□	□	□	□	□
3. I thought CyATP was easy to use.	□	□	□	□	□
4. I think that I would need the support of a technical person to be able to use CyATP.	□	□	□	□	□
5. I found the various functions in CyATP were well integrated.	□	□	□	□	□
6. I thought there was too much inconsistency in CyATP.	□	□	□	□	□
7. I would imagine that most people would learn to use CyATP very quickly.	□	□	□	□	□
8. I found CyATP very cumbersome to use.	□	□	□	□	□
9. I felt very confident using CyATP.	□	□	□	□	□
10. I needed to learn a lot of things before I could get going with CyATP.	□	□	□	□	□

Do you have any other comments, questions, or concerns? (Optional)

____

____

Questionnaire #2

Cybersecurity Awareness Training by CyATP

Serious Games (Quiz and Crossword Puzzle) Evaluation

For each of the following statements, please mark one box that best describes your reactions to CyATP today.

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
1. I was deeply concentrated on games.	□	□	□	□	□
2. If someone was talking to me, I couldn't hear him.	□	□	□	□	□
3. I forgot about time passing while using games.	□	□	□	□	□
4. I felt detached from the outside world while using games.	□	□	□	□	□
5. I think the games were fun.	□	□	□	□	□
6. I felt bored while using the games.	□	□	□	□	□
7. I really enjoyed studying with those games.	□	□	□	□	□
8. It felt good to successfully complete the tacks in those games.	□	□	□	□	□
9. I felt that those games can ease the way I learn.	□	□	□	□	□
10. Those games were a much easier way to learn compared to the usual teaching.	□	□	□	□	□
11. Those games made learning more interesting.	□	□	□	□	□
12. I felt that the games increased my knowledge.	□	□	□	□	□
13. I felt that I caught the basics of what I was taught with those games.	□	□	□	□	□

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
14. I will definitely try to apply the knowledge I learned with those games.	□	□	□	□	□
15. I received immediate feedback on my actions.	□	□	□	□	□
16. I could relate the content of those games to things I have done or thought about in real life.	□	□	□	□	□
17. It is clear to me how the content of games is related to things I already know.	□	□	□	□	□
18. The games' goals were presented clearly.	□	□	□	□	□
19. I think it was easy to learn how to use games.	□	□	□	□	□
20. I found the games unnecessarily complex.	□	□	□	□	□
21. I needed to learn a lot of things before I could get going with those games.	□	□	□	□	□
22. I felt that I needed help from someone else in order to use the games because It wasn't easy for me to understand how to control them.	□	□	□	□	□
23. It was easy for me to become skillful at using the games.	□	□	□	□	□
24. In some cases, there was so much information that it was hard to remember the important points.	□	□	□	□	□
25. I could not really understand quite a bit of the material in games.	□	□	□	□	□

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
26. The good organization of the content in concept map helped me to be confident that I would learn this material.	□	□	□	□	□
27. Those games did not hold my attention.	□	□	□	□	□
28. When playing the games, I did not have the impulse to learn more about the learning subject.	□	□	□	□	□
29. The games did not motivate me to learn.	□	□	□	□	□

Do you have any other comments, questions, or concerns? (Optional)

____

____

Questionnaire #3

Cybersecurity Awareness Training by CyATP

Learning Material Quality Evaluation
(Concept map / Concept text / Quiz / Crossword puzzle)
For each of the following statements, please mark one box that best describes your reactions to CyATP today.

		Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
1.	The cybersecurity knowledge covered by the concept map is adequate.	□	□	□	□	□
2.	The structure of the concept map for related keywords is suitable for learning.	□	□	□	□	□
3.	The concept text on the learn page is enough to understand related concepts.	□	□	□	□	□
4.	The length of the concept text on the learn page is appropriate.	□	□	□	□	□
5.	The questions in the quiz are grammatically correct.	□	□	□	□	□
6.	The meaning of the choices in the quiz is clear.	□	□	□	□	□
7.	The clues in the crossword puzzle can reflect the answers.	□	□	□	□	□
8.	Overall, I was satisfied with the quality of the learning material.	□	□	□	□	□

Do you have any other comments, questions, or concerns? (Optional)

____

____

Bibliography

[1] Ponemon Institute. *Cybersecurity in the Remote Work Era: A Global Risk Report.* 2020. URL: `https://www.keepersecurity.com/blog/2020/10/13/new-ponemon-report-shows-u-s-companies-struggling-with-remote-work-cybersecurity/`.

[2] IBM Security Services. *Cybersecurity Intelligence Index.* 2016. URL: `https://media.scmagazine.com/documents/82/ibm_cyber_security_intelligenc_20450.pdf`.

[3] atan. *What is CTF and how to get started!* 2019. URL: `https://dev.to/atan/what-is-ctf-and-how-to-get-started-3f04`.

[4] Jin-Ning Tioh, Mani Mina, and Douglas W Jacobson. "Cyber security training a survey of serious games in cyber security". In: *2017 IEEE Frontiers in Education Conference (FIE)*. IEEE. 2017, pp. 1–5.

[5] Tafadzwa Sigauke. "DESIGN AND IMPLEMENTATION OF A MOBILE BASED CHATBOT SYSTEM FOR SMALL SCALE MINERS USING ARTIFICIAL INTELLIGENCE". PhD thesis. Nov. 2019.

[6] Ehud Reiter and Robert Dale. *Building natural language generation systems*. Cambridge university press, 2000.

[7] AISmartz. *The Past and the Present of Natural Language Generation.* 2019. URL: `https://www.aismartz.com/blog/the-past-and-the-presence-of-natural-language-generation/`.

[8] Bidyut Das and Mukta Majumder. "Factual open cloze question generation for assessment of learner's knowledge". In: *International Journal of Educational Technology in Higher Education* 14.1 (2017), pp. 1–12.

[9] scikit-learn. URL: `https://scikit-learn.org/stable/`.

[10] Clark C Abt. "Serious Games". In: *New York City, New York, USA, st edition* (1970).

[11] Ben Sawyer and David Rejeski. *Serious games: Improving public policy through game-based learning and simulation.* 2002.

[12] Michael Zyda. "From visual simulation to virtual reality to games". In: *Computer* 38.9 (2005), pp. 25–32.

[13] Sebastian Deterding et al. “Gamification. using game-design elements in non-gaming contexts”. In: *CHI’11 extended abstracts on human factors in computing systems*. 2011, pp. 2425–2428.

[14] Kipp-report. *Power Play: Game-based training program launches in Middle East.* 2012. URL: `http://www.kippreport.com/fcs/power-play-game-based-trainingprogram-launches-in-middle-east/`.

[15] Asha Pandey. *Top 6 Benefits Of Gamification In eLearning*. 2015. URL: `https://elearningindustry.com/top-6-benefits-of-gamification-in-elearning`.

[16] Steve Sheng et al. “Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish”. In: *Proceedings of the 3rd symposium on Usable privacy and security*. 2007, pp. 88–99.

[17] Michael F Thompson and Cynthia E Irvine. “CyberCIEGE scenario design and implementation”. In: *2014 {USENIX} Summit on Gaming, Games, and Gamification in Security Education (3GSE 14)*. 2014.

[18] Fares Kayali et al. “A case study of a learning game about the Internet”. In: *International Conference on Serious Games*. Springer. 2014, pp. 47–58.

[19] Affan Yasin et al. “Improving software security awareness using a serious game”. In: *IET Software* 13.2 (2019), pp. 159–169.

[20] Hamed Alqahtani and Manolya Kavakli-Thorne. “Design and Evaluation of an Augmented Reality Game for Cybersecurity Awareness (CybAR)”. In: *Information* 11.2 (2020), p. 121.

[21] Dr Mark van Rijmenam. *7 Steps to Machine Learning: How to Prepare for an Automated Future*. 2019. URL: `https://medium.com/dataseries/7-steps-to-machine-learning-how-to-prepare-for-an-automated-future-78c7918cb35d`.

[22] Pranav Rajpurkar et al. “Squad: 100,000+ questions for machine comprehension of text”. In: *arXiv preprint arXiv:1606.05250* (2016).

[23] Zheyu Tan et al. “Adaptive security awareness training using linked open data datasets”. In: *Education and Information Technologies* 25 (2020), pp. 5235–5259.

[24] Matthew Honnibal et al. *spaCy: Industrial-strength Natural Language Processing in Python*. 2020. DOI: `10.5281/zenodo.1212303`. URL: `https://doi.org/10.5281/zenodo.1212303`.

[25] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. “Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning”. In: *Journal of Machine Learning Research* 18.17 (2017), pp. 1–5. URL: `http://jmlr.org/papers/v18/16-365.html`.

[26] Daiki Gyoten, Masato Ohkubo, and Yasushi Nagata. “Imbalanced data classification procedure based on SMOTE”. In: *Total Quality Science* 5.2 (2020), pp. 64–71.

[27] John Platt et al. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in large margin classifiers* 10.3 (1999), pp. 61–74.

[28] Bianca Zadrozny and Charles Elkan. "Transforming classifier scores into accurate multiclass probability estimates". In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* 2002, pp. 694–699.

[29] Alexandru Niculescu-Mizil and Rich Caruana. "Predicting good probabilities with supervised learning". In: *Proceedings of the 22nd international conference on Machine learning.* 2005, pp. 625–632.

[30] Sara Vincent. *Sentence length: why 25 words is our limit.* 2014. URL: `https://insidegovuk.blog.gov.uk/2014/08/04/sentence-length-why-25-words-is-our-limit/`.

[31] Horabail Venkatagiri. "Effect of sentence length and exposure on the intelligibility of synthesized speech". In: *Augmentative and Alternative Communication* 10.2 (1994), pp. 96–104.

[32] WebPageTest. URL: `https://www.webpagetest.org/`.

[33] J Brooke. *Usability evaluation in industry, chap. SUS: a "quick and dirty" usability scale.* 1996.

[34] Jeff Sauro. *A practical guide to the system usability scale: Background, benchmarks & best practices.* Measuring Usability LLC, 2011.

[35] Aaron Bangor, Philip Kortum, and James Miller. "Determining what individual SUS scores mean: Adding an adjective rating scale". In: *Journal of usability studies* 4.3 (2009), pp. 114–123.

[36] Emmanuel Fokides et al. "Let players evaluate serious games. Design and validation of the Serious Games Evaluation Scale". In: *ICGA Journal* 41.3 (2019), pp. 116–137.