Content Generation and Serious Game Implementation for Security Awareness Training

1910129 ZENG Youmeizi

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