Applied Science & Engineering Journal for Advanced Research *Peer Reviewed and Refereed Journal*
ISSN (Online): 2583-2468
Volume-3 Issue-3 || May 2024 || PP. 1-5 DOI: 10.5281/zenodo.11180356

Password Complexity Prediction Based on RoBERTa Algorithm

Yuhong Mo[1], Shaojie Li[2], Yushan Dong[3], Ziyi Zhu[4] and Zhenglin Li[5]
[1]*College of Engineering, Carnegie Mellon University, PA, Pittsburgh, 15213, USA*
[2]*Huacong Qingjiao Information Technology (Beijing) Co., Ltd., Beijing, China*
[3]*University of Maryland, MD, USA*
[4]*New York University, USA*
[5]*Texas A&M University, USA*

[1]*Corresponding Author: yuhongmo@cmu.edu*

Received: 09-04-2024 Revised: 25-04-2024 Accepted: 10-05-2024

ABSTRACT
Corresponding author email: In the digital age, password security is a top priority for protecting personal information. Machine learning techniques provide us with intelligent and efficient means to enhance password security. In this paper, we adopt RoBERTa algorithm and use the password complexity text dataset for password complexity prediction, and the confusion matrix and accuracy rate of the three classifications are derived through two model trainings. The confusion matrix shows that the vast majority of the classification results are accurate, and the accuracy of the two classifications is over 99.741% and 99.11%, respectively. This indicates that the model is able to effectively predict password complexity, provide users with accurate feedback, and prompt users to enhance password security in a timely manner. Through this study, we can better understand how to use machine learning technology to improve password security and protect personal private information from malicious intrusion. In our daily life, we should pay attention to the complexity of password settings and realise the importance of password security for personal information protection. We look forward to the launch of more similar studies in the future to further strengthen cybersecurity protection measures and work together to build a more secure and reliable digital environment.

Keywords: *password complexity, roberta, accuracy*

I. INTRODUCTION

Password security is crucial in today's digital society, which is directly related to personal privacy, property security, and information security. A strong password can effectively protect personal accounts from the threat of hacking, malware invasion, or information leakage [1]. The higher the complexity of a password, the more difficult it is to crack it, so setting and managing passwords appropriately is crucial for protecting personal information.

Machine learning plays an important role in predicting password security complexity. Machine learning algorithms can analyse a large amount of password data and mine the patterns and regularities in it to help users create more secure and complex passwords. For example, machine learning can identify common weak password patterns (e.g., consecutive numbers, simple dictionary words, etc.) and remind users to avoid these password combinations that can be easily guessed or cracked [2,3].

In addition, machine learning can generate personalised password suggestions based on users' personal information and usage habits. By analysing the user's preferences, birthdays, common vocabulary and other information, machine learning can recommend password combinations that meet personal preferences but have a certain degree of complexity, helping the user to strike a balance between memory convenience and security [4]. In addition, machine learning can be applied to detect password leakage and malicious attacks. By monitoring network traffic and login behaviour data, the machine learning system can detect abnormal activities in time and take corresponding measures to prevent hackers from using leaked passwords to carry out malicious login or vandalism.

Password security is the first line of defence for personal information security in the digital era, while machine learning technology provides us with more intelligent and efficient means to enhance password security [5]. Therefore, in our daily life we should pay attention to and strictly comply with the recommendations on password setting and management, and use technological means to continuously improve the safety and security level of our accounts and information.

ISSN (Online): 2583-2468
Volume-3 Issue-3 || May 2024 || PP. 1-5 DOI: 10.5281/zenodo.11180356

II. DATA SOURCES

The Password Complexity dataset is a dataset used to train and test password strength prediction models that contains the text of the passwords along with the corresponding password complexity labels (0, 1, and 2). This dataset is commonly used for password security research in the field of machine learning and is intended to help developers build algorithms that can automatically evaluate password strength. By using such a password complexity dataset, machine learning models can be trained to automatically recognise and assess the strength of user-created passwords and provide feedback and suggestions accordingly. This can help strengthen cybersecurity measures and raise users' awareness of personal information security.

The dataset used in this paper is an open source dataset, which classifies the complexity of passwords into 3 categories, the dataset consists of 2 columns, the first column is the passwords in the form of strings, and the second column is the complexity level of the passwords, whose complexity levels are 0, 1, and 2, where 0 is an unreliable password, 1 is a moderately reliable password, and 2 is a very reliable password. Part of the dataset is shown in Table 1, the password complexity of the data is counted and the result is shown in Figure 1. The length statistics of the texts of the three cipher complexities are carried out, i.e., each text is counted according to the number of words in the text, and finally the length information of each text is summarised, and the results are shown in Fig. 2.

Table 1: Partial data.

Password	Strength
yrtzuab476	1
yEdnN9jc1NgzkkBP	2
sarita99	1
Suramerica2015	2
PPRbMvDIxMQ19TMo	2
yuri2011	1
za1njutt	1
g4d96apen	1
amormio123	1
cat700700	1
7ZWdMXTI1MwJaAVo	2
maroccon123	1
9ke0bvq	0
3IJ2RUAB	1
124dscate	1
4DorrxDQ5OAI16PL	2
bentor123	1
kristian1997	1
rlh0u771	1
porseo74	1

ISSN (Online): 2583-2468
Volume-3 Issue-3 || May 2024 || PP. 1-5 DOI: 10.5281/zenodo.11180356

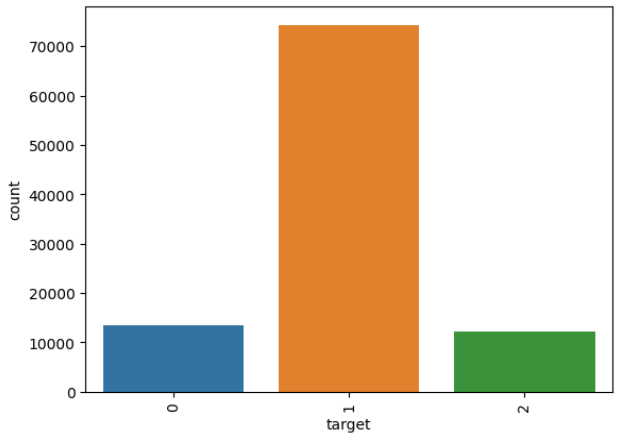

Figure 1: Statistical bar charts
（Photo credit : Original）

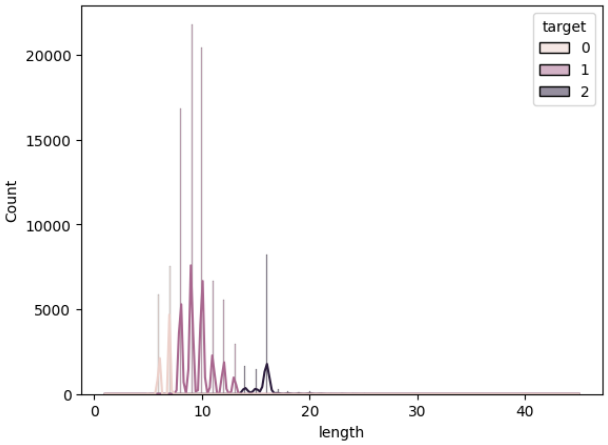

Figure 2: Statistical aggregation
（Photo credit : Original）

III. METHOD

RoBERTa (A Robustly Optimised BERT Approach) is a natural language processing model proposed by the Facebook AI team in 2019, which is based on the BERT (Bidirectional Encoder Representations from Transformers) model is improved and optimised [6,7].The principle of RoBERTa basically continues the architecture of BERT, but a series of improvements are made during the training and optimisation process, which makes RoBERTa achieve better performance on several natural language processing tasks.

RoBERTa uses a much larger dataset for pre-training. Compared to the original dataset used by BERT, RoBERTa uses more unlabelled text data and employs a higher number of iterations in the pre-training task, which helps the model learn the language representation better. In addition, RoBERTa introduces a dynamic masking strategy, i.e., randomly masking words at different positions in each training batch, which helps to improve the model's understanding of contextual information [8].

RoBERTa improves on the mask prediction task during pre-training. Unlike BERT, which uses the Next Sentence Prediction (NSP) task during pre-training to help the model understand the relationships between sentences, RoBERTa only uses the Masked Language Model (MLM) task, which means that it only needs to predict the correct form of the masked words [9]. This simplified pre-training objective allows the model to better learn the lexical representation and avoids the effects of noise that may be introduced by the NSP task.

In addition, RoBERTa employs a dynamic lexical replacement strategy to enhance the model generalisation capability. During the pre-training process, a portion of the masked vocabulary is randomly selected and replaced with other vocabulary, which is then restored back to the original vocabulary in a subsequent step [10]. This approach allows the model to

ISSN (Online): 2583-2468
Volume-3 Issue-3 || May 2024 || PP. 1-5 DOI: 10.5281/zenodo.11180356

have better adaptability to words that have appeared but not seen in the input data and improves its generalisation performance. In the fine-tuning phase, RoBERTa further optimises the performance by using larger scales, longer sequence lengths, and smaller learning rates. In addition, other techniques such as multi-task learning and migration learning can be combined during fine-tuning to further improve the model's performance on specific tasks.

Overall, by optimising the pre-training process, improving the mask prediction task, enhancing the generalisation capability, and optimising the fine-tuning phase, RoBERTa has achieved significant improvements on a variety of natural language processing tasks, and has become one of the most highly regarded models in the field today.

IV. EXPERIMENTS AND RESULTS

BATCH SIZE is set to 32, which divides the training set and test set in the ratio of 8:2, 80% of the data is used for training and 20% for testing. The SEQUENCE LENGTH is set to 60 for processing the length of the input text sequence. In defining the classifier, use the from preset method of RoBERTa classifier and pass the previously defined preprocessor as a parameter, specifying num classes as 3, i.e., the number of target classes is 3 classes. Sparse Categorical Crossentropy is used as the loss function, Adam optimiser is selected and the learning rate is set to 1e-5, using accuracy as the evaluation metric.

The model is trained twice to output the confusion matrix of the three classifications and their respective accuracies, and the confusion matrices of the two trainings are shown in Fig. 2 and Fig. 3, respectively.

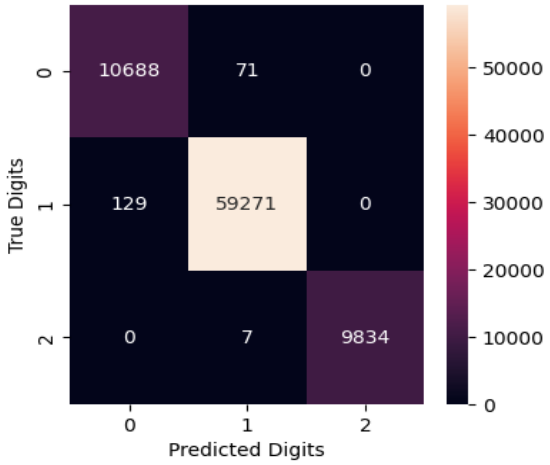

Figure 3: Confusion matrix
（Photo credit : Original）

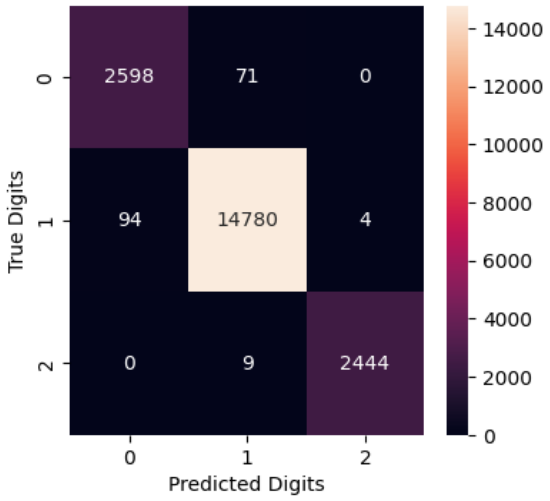

Figure 4: Confusion matrix
（Photo credit : Original）

ISSN (Online): 2583-2468
Volume-3 Issue-3 || May 2024 || PP. 1-5 DOI: 10.5281/zenodo.11180356

From the confusion matrix, it can be seen that the vast majority of classifications are accurate, and the accuracy of the two classifications is 99.741% and 99.11%, respectively, both of which are more than 99%, and the model is able to predict the complexity of passwords very well, give the user accurate feedback, and remind the user to increase the complexity of passwords in a timely manner.

V. CONCLUSION

The importance of password security as the first line of defence for personal information security in the digital age cannot be overstated. With the continuous evolution and intensification of cyber-attack techniques, simple and easy-to-crack passwords can no longer meet the needs of today's information security. Therefore, the use of machine learning techniques to enhance password security has become an innovative initiative. The method of password complexity prediction based on RoBERTa algorithm demonstrated in this paper not only enhances the intelligence level of password security detection, but also provides users with more convenient and efficient means of protection.

In this paper, the prediction of password complexity based on RoBERTa algorithm is successfully achieved by training the model twice on the basis of the selected textual dataset of password complexity, and the accurate prediction of password complexity is successfully achieved. The analysis of the confusion matrix shows that the vast majority of the classifications are accurate, and the accuracy of the two classifications reaches 99.741% and 99.11%, respectively, both exceeding the high accuracy level of 99%. This indicates that the constructed RoBERTa model performs well on the task of password complexity prediction, and is able to provide users with accurate feedback and timely reminders to increase the complexity of their passwords.

The experimental results in this paper show that password complexity prediction using the RoBERTa algorithm has a high degree of reliability and accuracy. an accuracy rate of more than 99% means that the model has a high degree of correctness in the prediction process, and is able to identify and assess password complexity effectively. This will greatly enhance users' alertness to personal information security issues and prompt them to take more rigorous and effective protection measures.

REFERENCES

1. Mu, Pengyu, Wenhao Zhang, & Yuhong Mo. (2021). Research on spatio-temporal patterns of traffic operation index hotspots based on big data mining technology. *Basic & Clinical Pharmacology & Toxicology, 128 (111), River ST, Hoboken*, 07030-5774, NJ USA: Wiley.
2. Mo, Y., Qin, H., Dong, Y., Zhu, Z., & Li, Z. (2024). Large language model (llm) ai text generation detection based on transformer deep learning algorithm. *International Journal of Engineering and Management Research*, *14*(2), 154-159.
3. Liu, B., Yu, L., Che, C., Lin, Q., Hu, H., & Zhao, X. (2023). *Integration and performance analysis of artificial intelligence and computer vision based on deep learning algorithms.* arXiv preprint arXiv:2312.12872.
4. Zhang, Jingyu, et al. (2024). *Research on detection of floating objects in river and lake based on AI intelligent image recognition.* arXiv preprint arXiv:2404.0688.
5. Xiang, Ao, et al. (2024). *Research on splicingimage detection algorithms based on natural image statistical characteristics.* arXiv preprint arXiv:2404.16296.
6. Dai, Shuying, et al. (2024). The cloud-based design of unmanned constant temperature food delivery trolley in the context of artificial intelligence. *Journal of Computer Technology and Applied Mathematics, 11*, 6-12.
7. Li, Zhenglin, et al. (2023). Stock market analysis and prediction using LSTM: A case study on technology stocks. *Innovations in Applied Engineering and Technology*, 1-6.
8. Li, Shaojie, Yuhong Mo, & Zhenglin Li. (2022). Automated pneumonia detection in chest x-ray images using deep learning model. *Innovations in Applied Engineering and Technology*, 1-6.
9. Mo, Y., Qin, H., Dong, Y., Zhu, Z., & Li, Z. (2024). Large language model (llm) ai text generation detection based on transformer deep learning algorithm. *International Journal of Engineering and Management Research, 14*(2), 154-159.
10. Mansouri, Mohamad, et al. (2023). Sok: Secure aggregation based on cryptographic schemes for federated learning. *Proceedings on Privacy Enhancing Technologies*.