International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 12 | December - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Password Strength Analysis: A Survey

Sukrut Pendharkar [1], Krutarth Bhatt [2], Sandeep Kumar [3], Prof. Saritha Murali [4]

[1, 2, 3, 4] *School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India*

Abstract - This paper outlines research based on password strength measurement analysis using various machine learning approaches. In the current world, a password is the most indispensable authentication mechanism concerning any system security.

Although there are multiple authentication methods available nowadays which provide improved security such as biometrics and smart cards still the password authentication mechanism is auspicious to everyone for enhanced system security. Generally, humans prefer to set a password related to their birthplace, date of birth, keyboard patterns, dictionary words, phone number, etc. Online or offline attackers and intruders can guess these passwords easily and intrude the system security. Privacy and data protection have become a need today. Cyber security is growing in its importance in literally every domain in recent years. Every person must safeguard their data and have a basic understanding of the technical developments taking place in many businesses.

Key Words: Machine learning, password, Multi-Layer Perceptron

I. INTRODUCTION

Most of the data (80%) available in social networks is unstructured. This makes it difficult to analyse and gain valuable judgment from such data. Sentiment analysis or opinion mining is an important technique, which helps in detecting the opinions of people on social media data. User Authentication is an important aspect of information security.

The most prevalent user authentication technique is passwords. Everyone who uses various online services is concerned about security and privacy to safeguard personal information from outsiders. Nowadays, life has become highly dependent on multiple security authentication systems. Passwords provide the first line of defense against unauthorized access to the personal information.

Amongst various kinds of authentication methods available [1] including bio-metric fingerprints, and security hardware tokens, password authentication has emerged as a strong authentication mechanism against various security attacks. In day-to-day life, various activities such as logging in to systems, websites, logging in to email accounts, fund transfer, online shopping, online exchanges, networks, databases, online trading, web portals, etc. require password authentication. But the main difficulties arise in selecting and utilizing highly secured passwords. Privacy and data protection have become a need today. Cyber security is growing in its importance in literally every domain in recent years. It is required for every individual to keep safe their data and have basic knowledge of what is happening in the industries in terms of technical aspects. One of the major concerns that people often have is to creating complex passwords. Their main fear is to forget them, particularly when there are several to remember. Naturally, a person tends to think of something that will be easy for them to memorize. One of the most common ways that hackers break into computers is by guessing passwords. Simple and commonly used passwords enable intruders to easily gain access and control of the device. Whereas a password that is difficult to guess makes it difficult for common hackers to break into a machine. The more difficult the password, the lower the likelihood that one's computer will fall victim to an unwanted intrusion.

In the proposed idea for Strength Analysis, the user would input their password and the system would analyse it using a machine Learning model and display if it is Weak or Strong with the accuracy percentages based on different strength levels.

Figure 1 shows the recent interest in "Password strength analysis" for the last 12 months according to Google Trends.

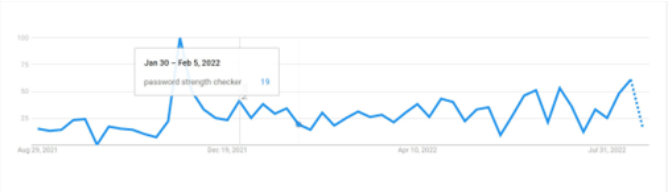

Fig.1. Recent interest in password strength analysis source: https://trends.google.com

II. PROBLEM STATEMENT

Everyone who uses various online services is concerned about security and privacy to safeguard personal information from outsiders. The password authentication system is one of several authentication systems available for the security of persons' data. We aim to develop a better password strength analyser using various machine learning algorithms to classify input passwords that could be both secure and easy to remember.

III. RELATED WORK

[1] describes the password strength checker based on the input strength levels defined in the dataset after which applying multiple machine learning models in order to find the better accurate result.

Authors Xinchun Cui, Xueqing Li, Yiming Qin, Ding Yong in paper [2] describes a technique to evaluate password strength by comparing user password to a password dictionary containing personal information to determine the match percentage.

Paper [3] shares the information about Machine Learning methods used for various consumer sentiment analysis and study. Strength analysis based on real life passwords using Markov Model [4] is a proactive password checker as a method for preventing users from creating easy-to-guess passwords. This model can be helpful in creating more effective password strength checker which will be able to track the probability percentage of given password entry chosen by the user.

[5] Tells a detailed review based on the mathematical aspects of Support Vector Machine model as a proactive password strength analyser based on certain filtering methods. Based on the results of the research, the Support Vector Machine (SVM) performs much better than any other machine learning algorithm used.

[6] Scikit learn is an open-source machine learning library which supports both supervised and unsupervised machine learning. It is a simple but yet efficient tool for predictive data analysis. It contains various classification, regression, clustering, preprocessing techniques which are useful in model fitting, predicting, cross-validation, etc.

A. System Architecture

The system architecture flow for the proposed review is shown in the below figure. It begins with the data collection stage, collecting the necessary raw data, and moves on to preprocessing, which removes noise from the data and extracts the useful information based on certain tools. After feature extraction, the data set is divided into two sets, one for training and the other for testing. Depending on the features and accuracy to be expected we can tweak the percentage of training and testing model. Generally, we apply 70-30 rule for training and testing purpose. After which model building will begin and results will be provided

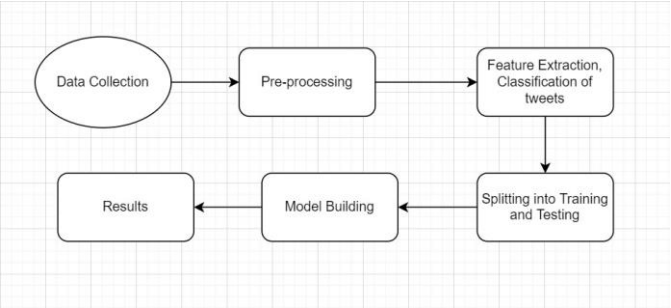

Fig. 2. System Architecture

Once the password entered by the user is categorized into any of the strength levels and if the output is either very weak, weak, or medium then the user can be prompted to enter his or her personal information like name, date of birth, place of birth, Mother's name, Father's name and from these provided user's personal information certain characters can be taken and along with the special characters a password can be generated which would not only satisfy the strength value of the password but also makes it easier for the user to remember the password.

B. Methodology

According to the gigantic research works that have been conducted regarding the analysis of the strength of passwords, huge numbers of password strength detection approaches have been identified. Out of which few algorithms are based on lexical analysis. Based on the results of research [4], the Support Vector Machine (SVM) performs much better than any other machine learning algorithm used.

The main objective of this proposed method is to classify the password based on the strength defined in levels by employing various machine learning techniques and models to compare the best accurate result. The working of the proposed idea is divided into phases viz data generation, collection and preprocessing, feature extraction, model building (training and testing phase), and evaluation. The system has following stages:

1) Input Data Acquisition: The raw data input is in the form of excel documents.
2) Preprocessing: The preprocessing Consists of a series of operations performed on the input data, which includes background noise reduction, empty data restoration, filtering etc. Various tools and techniques may be used for this purpose. The dataset will contain randomly generated password strings and target attributes as 0 or 1. 0 indicates the input tuple is not strong while 1 indicates the password is strong. Similarly, another dataset can contain certain levels defining the strength such as 0, 1, 2, and say 3 containing combinations of letters, numbers, and special characters in permutations. Figure 3 gives the dataset snap used for processing.

3) Feature Extraction: Feature extraction techniques are used to analyse the similarities between the text and are usually used in Natural password processing (NLP) as the algorithms cannot work on the raw text directly.
4) Training and testing phase: After the whole dataset is characterized and extracted, the next step is to train the machine learning model. Generally, 80% of the input data is given for training, and the rest 20% is used for testing purposes. Hence, multiple machine learning approaches can be tested to extract better accuracy to use the model for real-time data.
5) Machine learning classifiers: Machine learning approaches can learn domain-specific patterns from the text which leads to better classification results.

	A	B	C
1	index	password	strength
2	0	kzde5577	1
3	1	kino3434	1
4	2	visi7k1yr	1
5	3	megzy123	1
6	4	lamborghin1	1
7	5	AVYq1lDE4MgAZfNt	2
8	6	u6c8vhow	1
9	7	abcde	0
10	8	universe2908	1
11	9	as326159	1
12	10	asv5o9yu	1
13	11	612035180tok	1
14	12	jytifok873	1
15	13	WUt9lZzE0OQ7PkNE	2
16	14	jerusalem393	1
17	15	g067057895	1
18	16	52558000aaa	1
19	17	idofo673	1
20	18	6975038lp	1

Fig. 3. Sample Dataset

- Supervised Learning: Supervised approaches require labeled training documents, where the labels are generally the classes. There are four types of supervised classification approaches which are linear, probabilistic, rule-based, and decision tree.
- Unsupervised Learning: Unsupervised approaches in the field of sentiment analysis generally use clustering, which can classify data into different categories without specifying exactly which sentiment is represented by each category. Cluster analysis techniques can be categorized into Hierarchical and Partition clustering.
- Semi-Supervised Learning: Semi-Supervised Learning (SSL) approaches are also used when there are difficulties in obtaining labeled data, but unlike unsupervised approaches, this technique uses a small set of initial labeled training data to guide the feature learning procedure.
- Reinforcement Learning: Reinforcement Learning (RL) is a machine learning method where an agent is rewarded in the next time step based on the evaluation of its previous action. Reinforcement learning has been applied to solve different problems such as robot control, and it has been mostly used in games.

In this study, Sklearn (scikit-learn) library [6] which is widely used in python language for processing data is used. It features various machine learning libraries such as classification, regression, clustering, model selection, etc. The following machine learning classifiers can be employed for password strength classification into different categories.

- Logistic Regression
- Decision Tree
- Multi-Layer Perceptron (MLP)
- Random Forest
- Naïve Bayes
- Support Vector Machine (SVM)
- Adaboost etc.

According to the research work [1], the author has introduced various machine learning models including SVM, Decision Tree, Logistic Regression, Multilayer Perceptron, and Ada-boosting to check and compare the accuracy measures. Scikit learn is a python library that helps to process several data processing capabilities including classification, regression, clustering, and model selection.

C. Advantages

Because of the widespread usage of machine learning, the future is now more predictable. It enables firms and enterprises to keep an eye on consumer trends and behaviour and encourages the creation of new products. It is very easy to represent or visualize for extensive analysis of huge datasets. Information storage and retrieval is a very crucial part which can be made easy with help of learning models which are in turn time and cost efficient. Another very important advantage is it can reduce the inaccuracy of the system by removing the manual entries which can cause the errors.

IV. RESULTS AND DISCUSSION

This section details the implementation details and the results of the proposed method for analysing password strength. The first step is reading the input dataset. The first five entries from the dataset are shown in Figure 4.

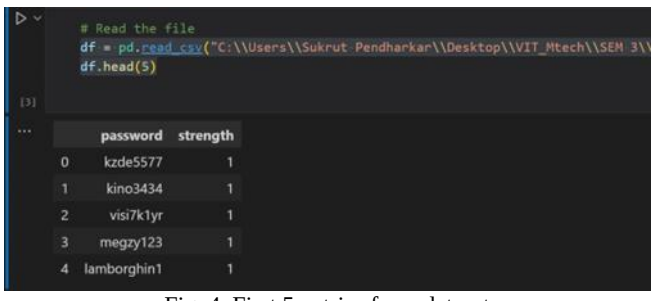

Fig. 4. First 5 entries from dataset

Counterplot based on the password strength values is given in Figure 5. After checking the count of input password values which contains only letters, numbers, combination of alphabets and numbers, combination of alphabets, numbers, and special characters, we applied various machine learning models including MLP in order to gauge the classification accuracy.

The dataset is then divided into training and testing phases as in Figure 6. Here, for our use case, we have distributed 80 percent for training and 20 percent for testing the data as shown below.

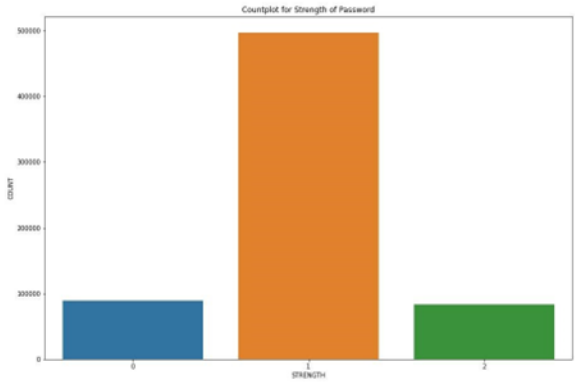

Fig. 5. Password strength

```
# Training and testing
xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size = 0.2, random_state = 42)
print(xtrain.shape, ytrain.shape)
print(xtest.shape, ytest.shape)

(535711, 5) (535711,)
(133928, 5) (133928,)
```

Fig. 6. Training and testing phase

Decision tree classification

Figure 7, 8 and 9 shows the final accuracy scores of Decision tree and Logistic regression respectively.

```
# Print a classification report
print(metrics.classification_report(y,predictions))
```

	precision	recall	f1-score	support
0	0.58	0.30	0.39	89260
1	0.84	0.94	0.89	494597
2	0.81	0.69	0.75	82797
accuracy			0.82	666654
macro avg	0.74	0.64	0.68	666654
weighted avg	0.80	0.82	0.80	666654

Fig. 7. Decision Tree

Fig. 8. Decision Tree Testing Accuracy

Logistic Regression

```
# Print a classification report
print(metrics.classification_report(y,predictions))
```

	precision	recall	f1-score	support
0	0.58	0.30	0.39	89260
1	0.84	0.94	0.89	494597
2	0.81	0.69	0.75	82797
accuracy			0.82	666654
macro avg	0.74	0.64	0.68	666654
weighted avg	0.80	0.82	0.80	666654

Fig. 9. Logistic Regression

D. Weka tool for visualization

Weka is a collection of machine learning algorithms for data mining tasks. It contains the tools for preprocessing, classification, regression, clustering, visualization etc. It provides many algorithms for machine learning tasks and are divided into several groups viz. bayes, functions, rules, trees etc. Following are the descriptions where we have used Decision Tree as well as Regression analysis where it shows the integral components such as the percentage of correlation coefficient, errors and the following graphs showing the outputs based on categories.

- Decision tree

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correlation coefficient                  0.1135
Mean absolute error                      0.3196
Root mean squared error                  0.5326
Relative absolute error                 99.6287 %
Root relative squared error             99.3533 %
Total Number of Instances              297
```

Fig. 10. Decision Tree Visualization in Weka

- Regression analysis

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correlation coefficient                  0
Mean absolute error                      0.3208
Root mean squared error                  0.536
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              297
```

Fig. 11. Regression Visualization in Weka

E. Visual representations of the password combination pattern

This Fig. 12 shows the count of password patterns with only small letters and their combinations.

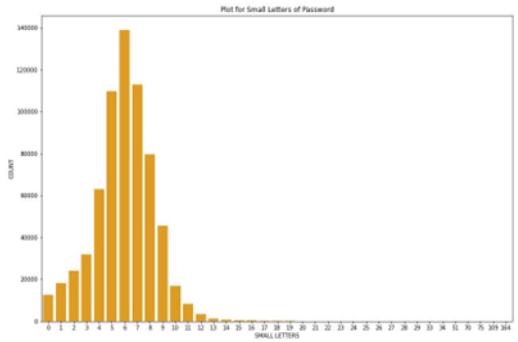

Fig. 12. Count of small letters password patterns

Figures 13 and 14 show the count of numeric password patterns and containing special characters combinations.

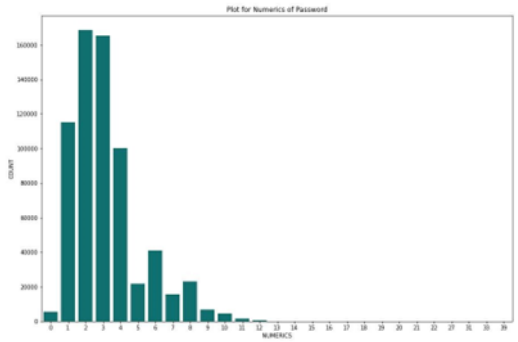

Fig. 13. Count of numeric password patterns

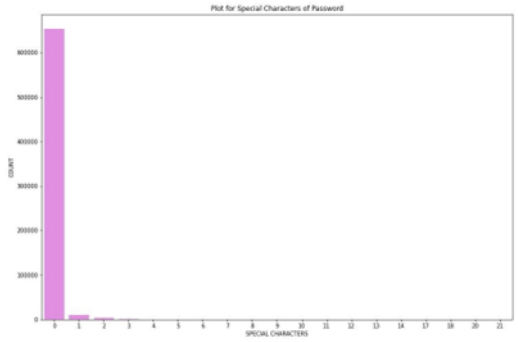

Fig. 14. Count of special characters password patterns

F. Multi-Layer Perceptron

MLP classifier stands for Multi-Layer Perceptron, which is linked to a Neural Network by its name. Unlike other classification techniques such as Support Vectors or Naive Bayes Classifier, this Classifier does classification using an underlying Neural Network. There are at least three levels of nodes in an MLP: an input layer, a hidden layer, and an output layer. Each node, with the exception of the input nodes, is a neuron with a nonlinear activation function. For training, MLP employs a supervised learning approach known as back propagation.

Multi-Layer Perceptron is distinguished from a linear perceptron by its numerous layers and non-linear activation. It can tell the difference between data that isn't linearly separable. A Multi-Layer Perceptron has input and output layers, and one or more hidden layers with many neurons stacked together. In each iteration, after the weighted sums are forwarded through all layers, the Mean Squared Error is computed across all input and output pairs. Then, if any error is found, and needs to be propagated back to the intermediate layer, the weights of the first hidden layer are updated with the value of the gradient.

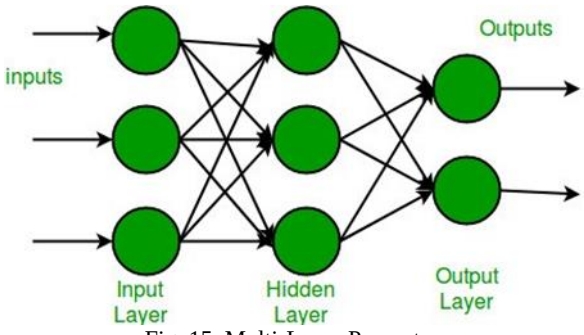

Fig. 15. Multi-Layer Perceptron

Scikit Learn library includes many machine learning algorithms including MLP where it contains the following parameters:

- Activation function: It decides whether a neuron should be activated on weighted sum and further addition of bias with it. Activation function decides whether a neuron should be activated or not by calculating weighted sum and further adding bias with it. In neural networks we update the weights on the basis of back propagation. When the function gets value greater than threshold it has a linear relationship with the dependent variables. There are various activation functions available such as Linear, RELU, Sigmoid, Tanh, RELU etc.
- Optimization function: Class of problems for finding the input to a given function that results in the minimum or maximum output from the function.
- Learning rate: It is a tuning parameter used in an optimization algorithm that determines the step size at each iteration. Usually, the minimum iterations given are 20 called "epochs".

A confusion matrix, also known as an error matrix, is used in the field of machine learning and specifically in the challenge of statistical classification. It is a specific table layout where each row represents the instances in an actual class and each column represents the instances in a predicted class, or vice versa. Figure 16 shows the Confusion matrix for our machine learning models used.

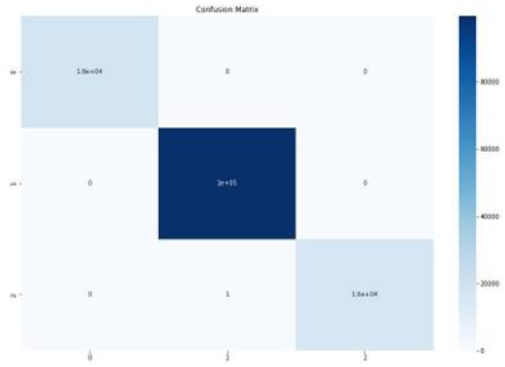

Fig. 16. Confusion matrix

V. CONCLUSIONS

Password authentication is required for a variety of activities in day-to-day life, including logging into systems and websites, logging into email accounts, fund transfers, and so on. Simple

VI. ACKNOWLEDGEMENT

With honesty we would like to express our deep sense of gratitude towards our esteemed guide Prof. Saritha Murali for giving us this splendid opportunity to select, research and present this project idea and also providing facilities for successful completion as well as priceless suggestions and for most valuable time lent as and when required.

VII. REFERENCES

[1] Sakya, Sakya Sarkar, and Mauparna Nandan Mauparna. ”Building a Multi-class Password Strength Generator and Classifier Model by Augmenting Supervised Machine Learning Techniques.” (2022).
[2] Cui, Xinchun, et al. ”A Password Strength Evaluation Algorithm Based on Sensitive Personal Information.” 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE, 2020.
[3] Jain, Praphula Kumar, Rajendra Pamula, and Gautam Srivastava. ”A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews.” Computer science review 41 (2021): 100413.
[4] Taneski, Viktor, et al. ”Strength Analysis of Real-Life Passwords Using Markov Models.” Applied Sciences 11.20 (2021): 9406.
[5] Suganya, G., S. Karpgavalli, and V. Christina. ”Proactive password strength analyzer using filters and machine learning techniques.” International Journal of Computer Applications 7.14 (2010): 1-5.
[6] https://scikit-learn.org/
[7] Dell’Amico, Matteo, Pietro Michiardi, and Yves Roudier. ”Password strength: An empirical analysis.” 2010 Proceedings IEEE INFOCOM. IEEE, 2010.
[8] Vijaya, M. S., K. S. Jamuna, and S. Karpagavalli. ”Password strength prediction using supervised machine learning techniques.” 2009 international conference on advances in computing, control, and telecommunication technologies. IEEE, 2009.
[9] Riley, Shannon. ”Password security: What users know and what they actually do.” Usability News 8.1 (2006): 2833-2836.