Porn Words Blocker Application using Aho-corasick Algorithm

Nicholas Chandra[1], Julio Christian Young[2], Seng Hansun[3]
[1]Informatics Department, Universitas Multimedia Nusantara, Indonesia, nicholas.chandra@student.umn.ac.id
[2]Informatics Department, Universitas Multimedia Nusantara, Indonesia, julio.christian@umn.ac.id
[3]Informatics Department, Universitas Multimedia Nusantara, Indonesia, hansun@umn.ac.id

ABSTRACT

Pornography is commonly used as a media for self-satisfaction called masturbation. Masturbation is considered as a harmful and addictive activity by a community called NoFap Indonesia. In this study, a custom keyboard application was built using the Aho-corasick algorithm to block words that are considered to have pornographic content that will be tested by members of NoFap Indonesia. Aho-Corasick algorithm is a dictionary-based algorithm that allows the fast transfer between strings that do not match on another branch which has the same prefix. This causes the system to move quickly without the need for backtracking or tracing. The usability test shows that the NFKeyboard application can be accepted, for the grade scale with B results which shows that the application has good value, and adjective ratings with excellent as a result. For the percentile rank gives a value of A. These results indicate that the NFKeyboard application has a perceived usability that is 90% higher than another product that has been tested. For the results of noting the word search speed measurement based on the index order and word length on two different smartphones' specifications, we concluded that the word length and word index order do not affect the speed results of the word search process.

Key words: Aho-corasick algorithm, Android, Custom keyboard, Pornography, System Usability Scale.

1. INTRODUCTION

With the Internet, benefits that we can get are numerous. Such as the ease of obtaining information, meeting others without having to meet, communicating in different places and times, and much more. However, it does not mean the Internet only contains benefits, there are also negatives. One of them is pornography.

The impact of pornography is very surprising, because turns out there are four effects that occur in those who are exposed to pornography. These effects include addiction, escalation, desensitization, and act out [1]. Furthermore, because they cannot apply act out into the real world, sexual desire is channeled in another way, namely masturbation. Based on the results of the distribution of questionnaires that have been conducted on the Indonesian NoFap community, exposed that addiction to pornography and masturbation harm themselves.

That is also why the test is being conducted on the Indonesian NoFap community.

The design and development were done by using Android Operating System because based on the results of research conducted by Juniper Research, during the first half of 2015 there were around 136 billion pornographic videos accessed via smartphones [2]. This application also uses the Aho-corasick algorithm, which was chosen because to find the appearance of multipattern in text strings, this algorithm is more appropriate because it performs exact matching of patterns in the text [3].

Because the algorithm is multipattern, there is no need to repeat the search process for each pattern. The text will not be read as many times as the number of patterns such as the single pattern algorithm, namely brute force, KMP, and Boyer-Moore [4]. The Aho-corasick algorithm is also the fastest string search algorithm [5].

Thus, the construction of a custom keyboard with the Aho-corasick algorithm can block pornographic words with a delay that is not too far between the process of blocking words and when the user typing the words.

2. THEORETICAL BASIS

2.1 Pornography

Pornography is a picture, sketch, illustration, photo, text, sound, moving picture, animation, cartoon, conversation, gestures, or other forms of message through various forms of communication media or public shows, which contain sexual obscenity or exploitation that is violating the norms of decency in society [6]. For those who have been exposed to pornography, there are four effects that will occur.

The effect consists of addiction, which is an effect where there is an addiction to see and get the material back. The second is escalation, where there is an increasing need for sex material that is heavier, more explicit, more sensational and more distorted than previously consumed. Third is desensitization which is the stage where sex material that was previously considered taboo, immoral, and demeaning human beings has now become something normal. Even become insensitive to victims of sexual violence.

Finally, there is an act out which is a tendency to apply to the real life of sexual behavior that has only been seen.

2.2 Aho-corasick

Aho-corasick is a multipattern algorithm that was invented by Alfred V. Aho and Margaret J. Corasick [7]. To find the emergence of multipattern in text strings, this algorithm is more appropriate because it performs exact matching of patterns in the text [3]. Because it is multipattern, there is no need to repeat the search process for each pattern.

The text will not be read as many times as the number of patterns such as the single pattern algorithm, namely brute force, KMP, and Boyer-Moore [4]. The Aho-corasick algorithm is also the fastest string search algorithm [5]. The following are the phases in Aho-Corasick:

1) Aho-corasick Preprocessing Phase
There are three steps in this preprocessing phase. The first is automata from which the pattern group {WOMAN, MAN, MEAT, and ANIMAL} is made of finite state automata (Figure 1).
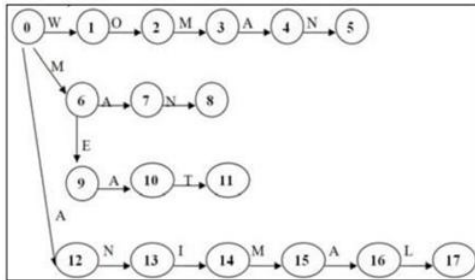

Figure 1: Automata [3]

After automata is created, then put it into the transition table. The automata transition can be seen in Figure 2.

Figure 2: Automata transition [3]

Second is the failure function (Figure 3), which means the longest suffix of a string which is also the prefix of several nodes.
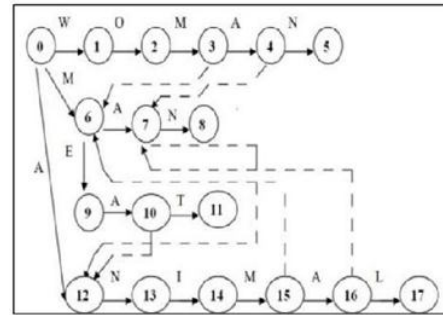

Figure 3: Failure Function [3]

Table 1: Failure Function Transition Table [3]

Node	Failure	Node	Failure
0	0	9	0
1	0	10	12
2	0	11	0
3	6	12	0
4	7	13	0
5	0	14	0
6	0	15	6
7	12	16	7
8	0	17	0

The purpose of this failure function (Table 1) is to enable the algorithm not to scan characters more than once [3]. Third is the output function, which is used to give a collection of patterns when entering the final state as shown in Figure 4 with transition table as shown in Table 2.
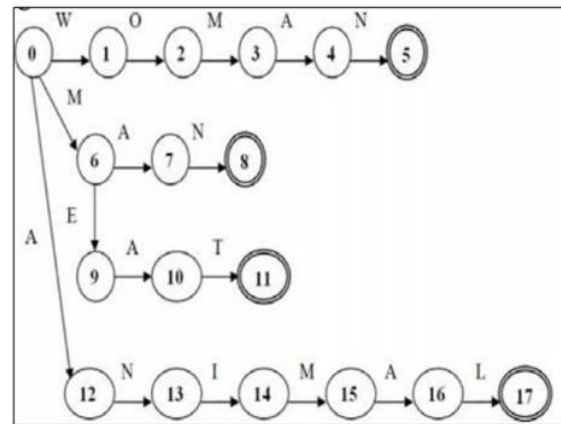

Figure 4: Output Function [3]

Table 2: Output Transition Table

Final State	Output
Node 5	WOMAN, MAN
Node 8	MAN
Node 11	MEAT
Node 17	ANIMAL

2)   *Aho-corasick Searching Phase*

To find text, Aho-Corasick uses a very easy way. When scanning text that is running through automata if a transition is found, do the transition, but if not found then check the failure function transition table.

2.3 System Usability Scale

System Usability Scale (SUS) is one of the user test methods that provides reliable "quick and dirty" measuring instruments [8]. The Usability Scale system is used because in testing it emphasizes the perspective of the end user so that the evaluation results will be more in line with the real situation [9].

According to research conducted by Tullis and Stetson, it shows that SUS produces the most reliable results than the Computer System Usability Questionnaire (CSUQ) and Questionnaire for User Interface Satisfaction (QUIS) [10]. Then according to research conducted by Sauro referred by Brooke shows that there are very similar results between those who try the system and then fill in the SUS questionnaire with a short time (5-6 seconds) and those who use the system for a longer time [11].

Based on research conducted by Lewis and Sauro in 2017, the results show that indeed the items from SUS are described as bidimensional items but the elaboration is not in a useful way. So, they recommend treating SUS as a unidimensional measure of perceived usability and no longer calculating or reporting sub-scales of the usability & learnability dimension [12].

System Usability Scale testing has 10 statements that are used as testing tools [13]. Table 3 shows the SUS questions.

Table 3: SUS questions [9]

| No.	Questions
1.	I think that I would like to use this system frequently
2.	I found the system unnecessarily complex
3.	I thought the system was easy to use
4.	I think that I would need the support of a technical person to be able to use this system
5.	I found the various functions in this system were well integrated
6.	I thought there was too much inconsistency in this system
7.	I would imagine that most people would learn to use this system very quickly
8.	I found the system very cumbersome to use
9.	I felt very confident using the system
10.	I needed to learn a lot of things before I could get going with this system

The assessment in SUS is as follows [8]: Scale strongly disagree (strongly disagree) to strongly agree (strongly agree) is worth 1 to 5. For odd statements: subtract 1 from the response given by the user. For even statements: 5 minus the response given by the user. Add the converted responses together and multiply the number by 2.5.

Then determine the average value of respondents' ratings [9]. Next is to determine the grade of the assessment results, which there are two ways that can be used [11].
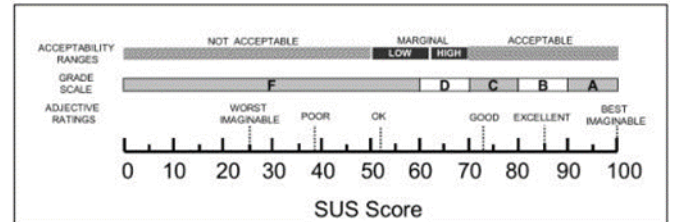
Figure 5: Grade range SUS Score [11]

The first determination is seen in terms of the acceptability ranges, grade scale and adjective rating. This first determination is obtained by comparing the results of the average total value with Figure 5. The second determination is seen in terms of the percentile range (SUS score) which can be seen in Figure 6.
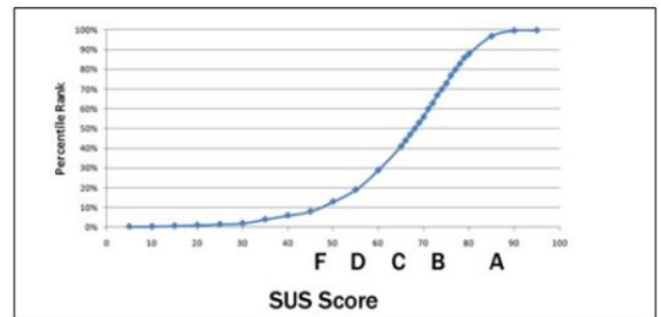
Figure 6: Percentile range SUS Score [11]

2.4 Simple Random Sampling

Simple random sampling is a sampling technique from members of the population that is done randomly without regard to strata that exist in that population. The terms of use of a simple random sampling technique are: First, this technique is used if the population element is homogeneous, so that any element selected as a sample can represent the population. Second, it is conducted if the research analysis tends to be descriptive and general in nature [14].
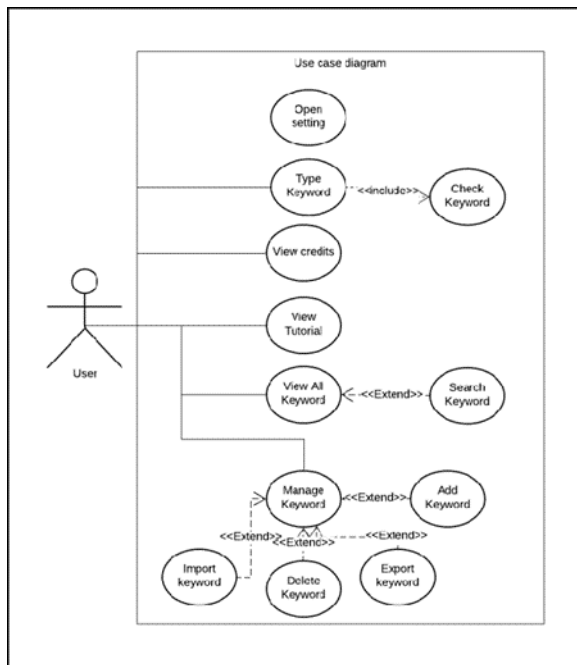
3. RESEARCH METHOD

3.1 Use Case Diagram

Figure 7: Use Case Diagram

In use case as shown in Figure 7, there is one actor, the user who uses the application. In the use case, the user is connected to 10 use cases, namely Add keywords, Search keywords, View all keyword keywords, Delete keywords, Type keywords, Tutorials, Import keywords, Export keywords, Open settings, and Credits. Use case Check keywords are included in the Use case Type keyword because when a user is typing sentences like normal typing on a keyboard made from this application, simultaneously Check keyword is also run.

Use case Add keyword, Delete keyword, Import keyword, and Export keyword are the extended use cases of the Use case Manage keyword because the user can do that, but that doesn't have to be done by the user. Likewise Search keywords also extended from the View all keyword because the user can do that but it does not have to be done by the user.

3.2 Methodology

a. Feasibility Studies
This Feasibility Study is conducted to find out whether it is true that NoFap Indonesia members are addicted to masturbation and pornography and to know whether the application to be made is indeed necessary or not. This measurement is done by distributing questionnaires with a minimum of 30 [15, 16].

b. Literature Studies
Collect and study relevant literature to achieve research results that are consistent with existing theories. In the study of literature learning of various theories related to the design and builder of android applications using Aho-corasick

algorithm. One theory taken for testing is the System usability scale.

c. System Design
Make an attractive interface, as well as the entire database related to application results. This can be done by making a diagram using Unified Modeling Language (UML). UML is used because it is commonly used as object-oriented modeling. UML itself consists of Use case diagrams, Activity diagrams, Sequence diagrams, and Class diagrams.

d. System Testing
This system testing phase is done when the system programming has been completed. Usability testing will be done with the System Usability Scale (SUS). Speed testing will be done by recording the time needed in the process of searching a word in the algorithm. The 500 words has been added to the database as keyword.

e. Evaluation
After testing the system, the evaluation is based on the results of the System Usability Scale and the results of testing the speed of the word search process.
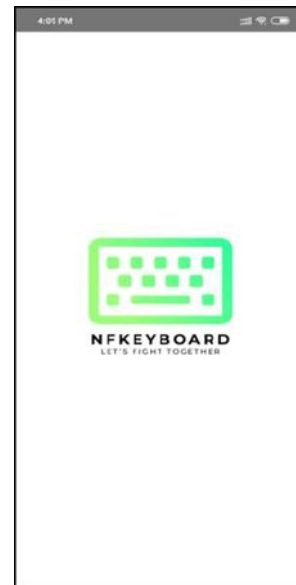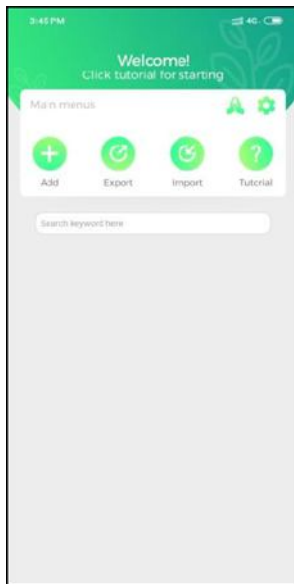
4. RESULTS AND ANALYSIS

4.1 Results

Figure 8: Splash Screen

When the user opens the application, it will immediately display the splash screen which can be seen in Figure 8. This splash screen will be displayed in one second. For that one second, user will be seeing the logo of the application that has been made.

Figure 9: Main menu

After splash screen is displayed, the application will immediately direct the user to the main page (Figure 9). This main page consists of add keyword button, export keyword button, import keyword button, tutorials button, credit button, and setting button. When the app is first installed, the main page looks blank or empty because the application does not have any keywords in the database. Which user has to add the keywords manually in the Add keyword menu.

Figure 10: Add Keyword menu

In this Add keyword menu that has been shown in Figure 10 users are expected to type in the fields provided then press the add button, so the application can add keywords to the database. As shown on Figure 11, can be seen the display of the add keyword page if the user succeeds in adding keywords. The keyword will also be directly displayed to the user on the main page.
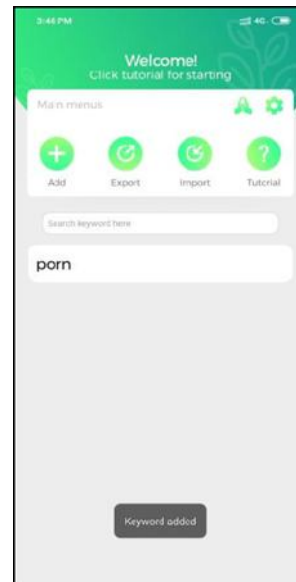
Figure 11: Keyword Added

Then in Figure 12, can be seen the NFKeyboard keyboard interface. The user must enable this keyboard in the open settings button first, then select NFKeyboard as the default keyboard in the smartphone settings so that it can be used. If the user does not enable NFKeyboard in their smartphone settings, NFKeyboard will not be usable or displayed when changing the keyboard to type. The keyboard interface that has been enabled and selected as the default keyboard can be seen in Figure 12.

Figure 12: NFKeyboard interface

Figure 13 shows a message on the NFKeyboard when the user types the keywords that have been added in the application. The word that has been added to the application will be immediately be blocked if the user types the keyword. The keyword will be deleted, but other than the keyword will be shown.

Figure 13: Keyword blocked

4.2 SUS Results

This section presents questionnaire results from 31 respondents using SUS. Raw score results for each respondent were obtained from the sum of each odd questions minus 1 and minus 5 the results of even questions [17]. Then SUS score is obtained from the raw score multiplied by 2.5 as shown in Table 4.

Table 4: System Usability Scale Results

No.	Raw Score	Sus Score	No.	Raw Score	Sus Score
1	35	87.5	16	21	52.5
2	35	87.5	17	37	92.5
3	35	87.5	18	35	87.5
4	40	100	19	34	85
5	27	67.5	20	33	82.5
6	35	87.5	21	34	85
7	23	57.5	22	40	100
8	37	92.5	23	36	90
9	35	87.5	24	26	65
10	33	82.5	25	36	90
11	34	85	26	35	87.5
12	38	95	27	29	72.5
13	36	90	28	22	55
14	31	77.5	29	35	87.5
15	34	85	30	31	77.5
			31	34	85
Total average			82.74193548		
Standard deviation			12.01		

To calculate the speed of the algorithm, first the words are taken from the 500 words that have been entered into the database. After that the keywords divided into 5 ranges. The ranges are 1-100, 101-200, 201-300, 301-400, 401-500. Then from each range the longest and shortest words are taken, and each taken word is tested 10 times. The results of the

recapitulation of the total calculation words are shown in Table 5 and Table 6 respectively.

Table 5: Shortest words calculation

Shortest Words					
	xnxx	Fux	Ass	Feet	vJav
A	0.486	0.511	0.511	0.554	0.548
B	1.934	1.953	1.640	1.817	1.888

Table 6: Longest words calculation

Longest Words					
	xMovies for You	Collection of Best Porn	Bottomless_ Vixens	Hot Chicks with Tattoos	Vintage-Erotica-Forum
A	0.526	0.563	0.536	0.572	0.602
B	1.961	1.803	1.768	1.901	1.994

5. CONCLUSION

From the research results, pornographic word blocker keyboard application on android with the Aho-corasick algorithm was successfully designed and built using the Java programming language and the Android Studio IDE. Based on the questionnaire data compiled based on SUS, the results of calculation of the data obtained were 82.74194 and obtained a standard deviation of 12.01 out of 100, which means that the data distribution of each respondent was said to be good.

From the results of the SUS score, it was obtained that the Acceptability ranges with acceptable results showed that the NFKeyboard application was acceptable, The Grade scale with the B results showed that the application had good value, and the Adjective ratings with excellent results. The result of percentile rank is a value of A. These percentile rank result indicate that the NFKeyboard application has a perceived usability that is 90% higher than the product that has been tested.

Based on data recording measurements of word search speed based on the index order and word length on two different smartphone specifications, it was concluded that the word length and word order do not affect the speed results of the word search process. The implementation of more advanced methods, such as ANN [18] and SVM [19] can be done for the future research.

ACKNOWLEDGEMENT

REFERENCES

1. E. Supriati and S. Fikawati. **Efek Paparan Pornografi pada Remaja SMP Negeri Kota Pontianak Tahun 2008**, *MAKARA, SOSIAL HUMANIORA*, vol.13, no.1, 2009: 48.
   https://doi.org/10.7454/mssh.v13i1.210

2. R. Surahman, Riset: Ada 136 Miliar Video Porno yang Ditonton Via Smartphone – ENCIETY NEWS. Available at
   https://www.enciety.co/riset-ada-136-miliar-video-porno-yang-ditonton-via-smartphone/.

3. S. Hasib, M. Motwani, and A. Saxena. **Importance of Aho-Corasick String Matching Algorithm in Real World Applications**, *International Journal of Computer Science and Information Technologies*, vol. 4, no. 3, pp.467–469, 2013.

4. V. N. Sahara, *Studi Perbandingan dan Implementasi Algoritma Aho-corasic String Matching dengan Algoritma Interpolation Search pada Aplikasi Kamus Kedokteran Berbasis Mobile*, Thesis, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, 2014.

5. S. Shanivan, N. Safriadi, and Yulianti. **Rancang Bangun Aplikasi Game Edukasi Shiritori Bahasa Indonesia dengan Algoritma Ahocorasick Berbasis Web**, *Jurnal Sistem dan Teknologi Informasi*, vol.5, no.1, pp.1-5, 2017.

6. Kementrian Agama, Undang-Undang Republik Indonesia Nomor 44 Tahun 2008 Tentang Pornografi, 2008.

7. A.V. Aho and M.J. Corasick, **Efficient String Matching: An Aid to Bibliographic Search**, *Communication of ACM*, vol.18, no.6, pp.333-340, 1975.
   https://doi.org/10.1145/360825.360855

8. A.W. Soejono, A. Setyanto, and A.F. Sofyan. **Evaluasi Usability Website UNRIYO Menggunakan System Usability Scale (Studi Kasus : Website UNRIYO)**, *Jurnal Teknologi Informasi*, vol.XIII, no.1, pp.29-37, 2018.

9. U. Ependi, F. Panjaitan, and Hutrianto. **System Usability Scale Antarmuka Palembang Guide Sebagai Media Pendukung Asian Games XVIII**, *Journal of Information Systems Engineering and Business Intelligence*, vol.3, no.2, pp.101-107, 2017.
   https://doi.org/10.20473/jisebi.3.2.80-86

10. T.S. Tullis and J.N. Stetson, **A Comparison of Questionnaires for Assessing Website Usability,** *UPA 2004 Presentation*, pp.1-12, 2004.

11. J. Brooke. **SUS: A Retrospective**, *Journal of Usability Studies*, vol.8, no.2, pp.29-40, 2013.

12. J.R. Lewis and J. Sauro, **Revisiting the Factor Structure of the System Usability Scale**, *Journal of Usability Studies*, vol.12, no.4, pp.183-192, 2017.

13. J. Brooke, SUS - A Quick and Dirty Usability Scale, 1996. Available at
    https://hell.meiert.org/core/pdf/sus.pdf.

14. A. Hidayat, Pengertian Simple Random Sampling, Jenis dan Contoh, 2018. Available at
    https://www.statistikian.com/2018/02/pengertian-simple-random-sampling.html.

15. I. Yunita and S. Hansun. **Automatic News Blog Classifier Using Improved K-Nearest Neighbor and Term Frequency-Inverse Document Frequency**, *Journal of Theoretical and Applied Information Technology*, vol.97, no.15, 2019, pp.4202-4212.

16. K. Dhammayanti, A. Wicaksana, and S. Hansun. **Position Placement DSS Using Profile Matching and Analytical Hierarchy Process**, *International Journal of Scientific & Technology Research*, vol.8, no.11, 2019, pp.204-207.

17. E. Sudarmilah and M. Thoyib. **Stop Demam Berdarah Dengue (DBD) – Virtual Reality Application for Learning about Dengue Fever**, *International Journal of Emerging Trends in Engineering Research*, vol.7, no.11, 2019, pp.441-448.
    https://doi.org/10.30534/ijeter/2019/077112019

18. A.V. Lakshmi, K.V.T.N. Babu, M.S.R. Deepak, A.S. Kumar, G.V.P.C.S. Yadav, V.G. Tilak, and V.S. Ghali. **A Machine Learning based Approach for Defect Detection and Characterization in Non-Linear Frequency Modulated Thermal Wave Imaging**, *International Journal of Emerging Trends in Engineering Research*, vol.7, no.11, 2019, pp.517-522.
    https://doi.org/10.30534/ijeter/2019/197112019

19. N.C.S. Reddy, P.C.R. Vemuri, and A. Govardhan. **An Emperical Study on Support Vector Machines for Intrusion Detection**, *International Journal of Emerging Trends in Engineering Research*, vol.7, no.10, 2019, pp.383-387.
    https://doi.org/10.30534/ijeter/2019/037102019