JOIN (Jurnal Online Informatika)
Volume 3 No. 2 | December 2018: 99-102
DOI: 10.15575/join.v3i2.196

Detect Malware in Portable Document Format Files (PDF) Using Support Vector Machine and Random Decision Forest

Abdachul Charim[1], Setio Basuki[2], Denar Regata Akbi[3]
[1, 2, 3]Informatics Department, Universitas Muhammadiyah Malang, Malang, Indonesia
[1]abdachulcharim@gmail.com, [2]setio_basuki2@yahoo.co.id, [3]dnarregata@umm.ac.id

Abstract-Portable Document Format is a very powerful type of file to spread malware because it is needed by many people, this makes PDF malware not to be taken lightly. PDF files that have been embedded with malware can be Javascript, URL access, media that has been infected with malware, etc. With a variety of preventive measures can help to spread, for example in this study using the classification method between dangerous files or not. Two classification methods that have the highest accuracy value based on previous research are Support Vector Machine and Random Forest. There are 500 datasets consisting of 2 classes, namely malicious and not malicious and 21 malicious PDF features as material for the classification process. Based on the calculation of Confusion Matrix as a comparison of the results of the classification of the two methods, the results show that the Random Forest method has better results than Support Vector Machine even though its value is still not perfect.

Keywords- portable document format*, malware, classification, *support vector machine, random forest

I. INTRODUCTION

The number of PDF features and supported by various devices and platforms so that this can be used by parties who are not responsible for spreading malware. Based on statistics from www.virustotal.com, PDF ranks first in document files infected with malware compared to other document files whereas when compared to file formats other than document file format, PDF is ranked third[1]. As explained by the picture in figure 1.

The reason for choosing these two algorithms is because they have a higher level of accuracy than others[3][4]. Support Vector Machine is an algorithm with the technique of determining the best hyperplanes to separate between classes of data that have been determined[5][6], while Random decision forest is an algorithm by applying several Decision Tree to vote for features as data classification[7].

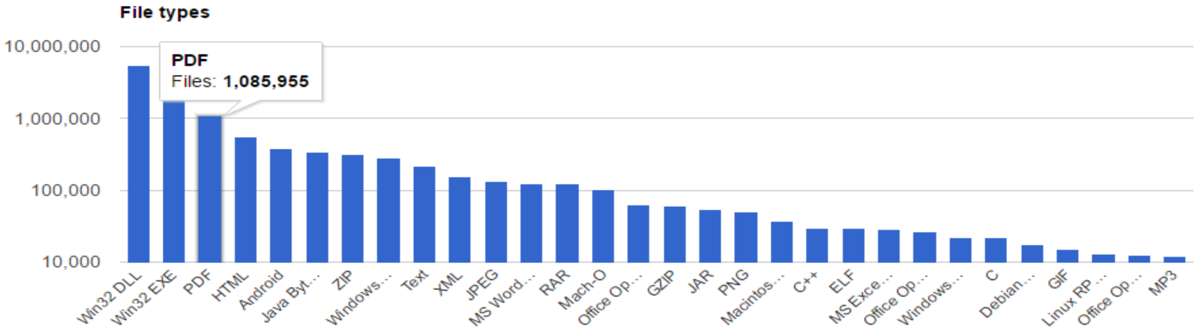


Figure 1. Results of Statistics www.virustotal.com[1]

PDF files that have been infected with malware can directly interfere with the performance of a system if the file is run. The impact of this is often not realized by ordinary users. To be able to distinguish between PDF files that have been infected with malware and those that cannot be done by analyzing malware detection using Machine Learning who have been trained to use a sample of malware that already exists and has been previously recognized[2]. Machine Learning has several algorithms that can be used, in this study two algorithms are used to compare the percentage of accuracy in the detection of PDF malware, namely Support Vector Machine and Random Forest.

The PDF file format has its hierarchical structure, by observing the structure can be identified if there is a dangerous code that has been implanted by irresponsible parties. From within the hierarchy structure features can be made as variables in determining whether the file is infected with malware or not. With the 2 classification methods above these features are used as classification variables between Malware and Non Malware classes.

Received: December 3, 2018; Revised: January 25, 2018; Accepted: January 25, 2019; Published: January, 31 2019

II. METHOD

The first thing to do in this research is collecting data in the form of 500 pdf files taken from various sources, from all the pdf files it is known whether or not the virus is infected because it is only for research purposes. After the pdf file is obtained, the file extraction process is made into a dataset which is then divided into training data and test data with a comparison of 350 training data and 150 test data. The classification process uses two methods, namely Support Vector Machine and Random Forest with pdf files and the number of comparison of training data and the same test data. From the results obtained in the two classification methods, it can be calculated using the Confusion Matrix to conclude which of the two methods is better as a classification of pdf files. The details of the research steps showed on figure 2.

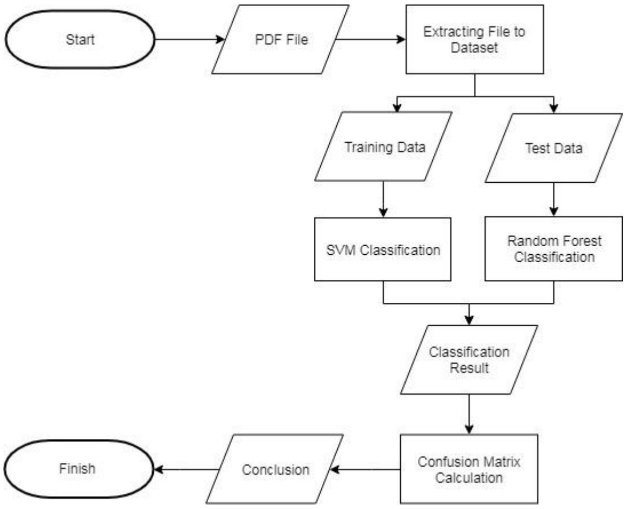
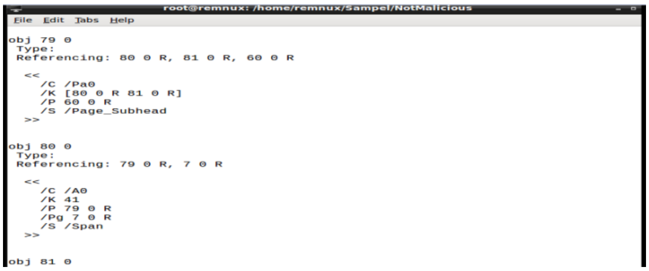


Figure 2. The flow of Research Methods

A. Portable Document Format (PDF)

PDF is a document file that is widely used on various devices and platforms because of its ease of use and can Include multimedia files, direct URL access, and HTTP communication[8]. In PDF files there is a hierarchical structure that consists of four elements, namely Objects, File Structure, Document Structure, and Content Streams. In Figure 3 can be explained on the left side is the physical layout of the PDF structure, in the middle is the logical structure, and on the right is the number of sets of structural paths. Below is an example of extraction from the PDF structure based on PDF Structure Representation.



```
obj 79 0
 Type:
 Referencing: 80 0 R, 81 0 R, 60 0 R

  <<
    /C /Pa0
    /K [80 0 R 81 0 R]
    /P 60 0 R
    /S /Page_Subhead
  >>


obj 80 0
 Type:
 Referencing: 79 0 R, 7 0 R

  <<
    /C /A0
    /K 41
    /P 79 0 R
    /Pg 7 0 R
    /S /Span
  >>


obj 81 0
```

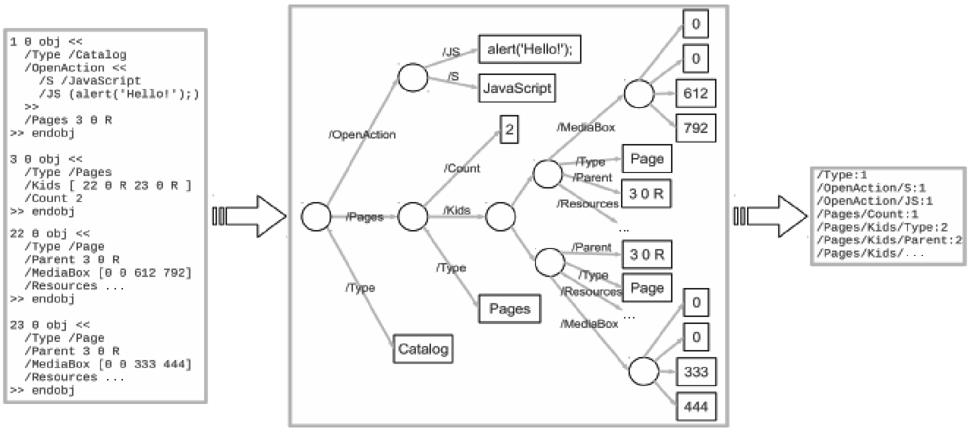
Figure 4. Results of PDF Metadata Extraction

B. Malicious PDF

Along with the use of many PDF format files, there is also a large spread of malware through PDF. PDF files that have been infected with malware can directly interfere with the performance of a system if the file is run. The impact of this is often not realized by ordinary users. With several factors above, it cannot be considered trivial so that spread can be prevented[9].

C. PDF Malicious Feature

In this study needed features that support identifying PDF malware files. This feature is obtained from extraction using the Pdfid.py program. These features are shared between 2 groups, including (1) Based on complete structure: Obj, EndObj, Stream, Endstream, Xref, StartXref, and Trailer. (2) Based on malicious PDF features: Page, Encrypt, ObjStm, JS, JavaScript, AA, OpenAction, Profile, JBIG2Decode, RichMedia, Launch, Embedded File, XFA, and Colors> 224[10].



```
1 0 obj <<
  /Type /Catalog
  /OpenAction <<
    /S /JavaScript
    /JS (alert('Hello!');)
  >>
  /Pages 3 0 R
>> endobj

3 0 obj <<
  /Type /Pages
  /Kids [ 22 0 R 23 0 R ]
  /Count 2
>> endobj

22 0 obj <<
  /Type /Page
  /Parent 3 0 R
  /MediaBox [0 0 612 792]
  /Resources ...
>> endobj

23 0 obj <<
  /Type /Page
  /Parent 3 0 R
  /MediaBox [0 0 333 444]
  /Resources ...
>> endobj
```

```
/Type:1
/OpenAction/S:1
/OpenAction/JS:1
/Pages/Count:1
/Pages/Kids/Type:2
/Pages/Kids/Parent:2
/Pages/Kids/...
```

Figure 3. PDF Structure Representation

D. Classification

In the identification of PDF malware files in this study using the file classification method based on PDF malware features. The classification process also requires training data and test data that has a total of 500 data. The classification method used is two types, namely Support Vector Machine and Random Forest.

E. Test Method

After the classification process from the Support Vector Machine and Random Forest methods, the classification results from files that are embedded with malware and not. After that, accuracy is sought for each classification. The calculation formula used to calculate the accuracy of the classification results using Confusion Matrix[11] which consists of:

- Accuracy (AC)

  $AC = (a + d) / (a + b + c + d)$
- Recall or True positive rate (TP)

  $TP = d / (c + d)$
- False positive rate (FP)

  $FP = b / (a + b)$
- True negative rate (TN)

  $TN = a / (a + b)$
- False negative rate (FN)

  $FN = c / (c + d)$
- Precision (P)

  $P = d / (b + d)$

Below is a table that explains the value of the formulas above:

Table 1. Confusion Matrix Tables

		Prediction	
		Negative	Positive
Actual	Negative	***a***	***b***
	Positive	***c***	***d***

- a is the true predictive value of a negative value
- b is the wrong predictive value of a positive value
- c is the wrong predictive value of a negative value
- d is the true predictive value of a positive value

III. RESULTS AND DISCUSSION

In the classification process using a dataset of 500 data, and divided into 2 class categories totaling 250 Malicious data and 250 data Not Malicious. After the classification process, the value of the confusion matrix is then calculated to get the Accuracy value, Recall / True Positive Rate, False Positive Rate, True Negative Rate, False Negative Rate, and Precision. Below is the result of the calculation:

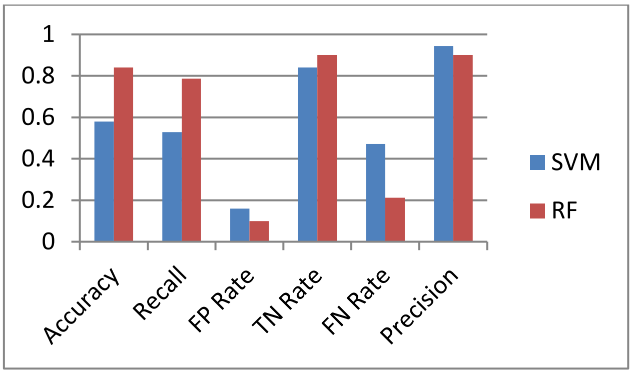


Figure 5. Results of Comparison in the Form of Graphs

Table 2. Comparison Results in Table Forms

	Accuracy	Recall/ TP Rate	FP Rate	TN Rate	FN Rate	Precision
SVM	0,58	0,528	0,16	0,84	0,472	0,943
Random Forest	0,84	0,787	0,1	0,9	0,212	0,9

By using the Confusion Matrix to compare the accuracy between the SVM and Random Forest classification methods, we get a Random Forest that has a higher Accuracy, Recall, and TN Rate by keeping the FP Rate and FN Rate lower even though the Precision value is slightly lower than SVM. So Random Forest is more accurate in the PDF Malware classification process using the features above.

IV. CONCLUSIONS

In this study focusing on the comparison of accuracy between the two classification algorithms using the malicious PDF feature, and the results obtained have a significant difference in the value of accuracy and recall. Random Forest has better results than Support Vector Machine when used as a classification method using this malicious PDF feature.

With the existence of deficiencies in this study development can be carried out with the following suggestions:

1. Increase the training dataset so that the system gets more training for classification.
2. The process of making datasets automatically with input files in .pdf format.
3. By using a classification method other than that used to find out how accurate the method is for classification of malware pdf files.

V. REFERENCES

[1] V. Total, “File Statistics.” [Online]. Available: https://www.virustotal.com/en/statistics/. [Accessed: 23-Jan-2018].

[2] J. S. Cross and M. A. Munson, “Deep PDF Parsing to Extract Features for Detecting Embedded Malware,” 2011.

[3] C. Smutz and A. Stavrou, “Malicious PDF detection using metadata and structural features,” in *ACSAC ’12 Proceedings of the 28th Annual Computer Security Applications Conference*, pp. 239–248.

[4] N. Šrndic and P. Laskov, “Detection of malicious pdf files based on hierarchical document structure,” in *In Proceedings of the Network and Distributed System Security Symposium, NDSS 2013*, 2012.

[5] K. Sembiring, *Penerapan Teknik Support Vector Machine untuk Pendeteksian Intrusi pada Jaringan*. Institut Teknologi Bandung, 2007.

[6] “Support Vector Machines - Scholastic Video Book Series,” *Scholastic Tutors*, 2014. [Online]. Available: https://scholastictutors.webs.com/Scholastic-Book-SupportVectorM-Part01-2014-01-26.pdf.

[7] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[8] A. Acrobat, “What is PDF?” [Online]. Available: https://acrobat.adobe.com/sea/en/acrobat/about-adobe-pdf.html?promoid=CW7625ZK&mv=other. [Accessed: 11-Feb-2017].

[9] D. Stevens, “Malicious PDF documents explained,” *IEEE Secur. Priv.*, vol. 9, no. 1, pp. 80–82, 2011.

[10] L. Rocha, “Malicious Documents - PDF Analysis in 5 Steps,” 2014. [Online]. Available: https://countuponsecurity.com/2014/09/22/malicious-documents-PDF-analysis-in-5-steps/. [Accessed: 12-Feb-2017].

[11] R. Kohavi and F. Provost, “Confusion matrix,” *Mach. Learn.*, vol. 30, no. 2–3, pp. 271–274, 1998.