Malicious Document Detection and Robust ML Model Construction

Yubin Yanga,\* , Wei Jiangb, Fengjiao Wangb ,Shumin Wei d

aSchool of Computer Science & Engineering, South China University of Technology, 510641 Guangzhou, China

bBluedon AI Lab, 510641 Guangzhou, China

cBluedon AI Lab, 510641 Guangzhou, China

dBluedon AI Lab, 510641 Guangzhou, China

Abstract

With the rapid development of information technology, it has become increasingly more important to perform detection on malicious documents. However, due to the diversity of document structures, attackers have gradually acquired a large attack vector. In this paper, we aim to construct a robust artificial intelligence (AI) document classifier both for industry and academia. Approximately 200,000 samples have been collected, and the AI model has been trained and optimized. The experimental results show that the accuracy of the model is as high as 99.82%, while the false-positive rate is as low as only 0.01%. Moreover, through the study of adversarial machine learning, the model is capable to resist attacks and enjoys good robustness. Finally, we demonstrate that our model can be widely deployed in typical usage scenarios, such as security products or mail servers.

*Keywords*: AI Security; Machine Learning; Maldoc Detection; Adversarial ML

© 2018 Totem Publisher, Inc. All rights reserved.

1. Introduction

Cyber attackers are turning to document-based malware as suggested by many anti-virus (AV) vendors. Users are increasingly being warned more generally of the danger of executable files by browsers, email agents, or AV products, but documents such as PDFs are treated with much less caution and scrutiny because of the impression that they are static files and can do little harm.

However, over time, PDF specifications have changed. The added scripting capability makes it possible for documents to work in almost the same way as executable files, including their ability to connect to the Internet, run processes, and interact with other programs. The growth of content complexity gives attackers more weapons with which to launch powerful attacks and more flexibility to hide malicious and evade detection.

A maldoc usually exploits one or more vulnerabilities in its interpreter to launch an attack. Unfortunately, given the increasing complexity of document readers and the wide library of component dependencies, attackers are presented with a large attack surface. New vulnerabilities continue to be found, with 137 published CVEs in 2015 and 227 in 2016 for Adobe Acrobat Reader (AAR) alone. The popularity of AAR and its large attack surface make it among the top targets for attackers. The collected malware samples show that many Adobe components have been exploited, including element parsers and decoders, font managers, and the JavaScript engine.

The continued exploitation of AAR along with the ubiquity of the PDF format makes maldoc detection a pressing problem, and many solutions have been proposed in recent years to detect documents bearing malicious payloads. These techniques can be classified into two broad categories: static and dynamic analysis.

Static analysis, or signature-based detection, parses the document and searches for indications of malicious content, such as shellcode or similarity to known malware samples. Dynamic analysis, or execution-based detection, runs the partial or entire document and traces malicious behaviors, such as vulnerable application programming interface (API) calls or return-oriented programming (ROP).

In the first half of this paper, we utilize machine learning techniques on document-specific attributes to identify embedded malware. Our approach addresses some of the shortcomings of existing techniques by use of a broadly applicable mechanism to classify and characterize documents.

As part of our analysis, we show that while the use of documents as an exploitation vector can be an enabling mechanism for the attacker, it also provides additional detection opportunities. All of the data closely associated with malicious activities can be used to aid detection, regardless of whether the data utilized for detection are inherently malicious or not. The underlying premise and intuition of our study are that malicious documents do have similarities to other malicious documents; they also have dissimilarities to benign documents, regardless of the specific vulnerability exploited or the specific malware embedded in the document. We posit that features based on document structure and metadata are adequate for reliable document classification given that appropriate statistical methods are applied to these features. This ensemble classifier is also able to classify previously unseen variants.

Clearly, deployment of learning methods in any security-critical context requires that they can withstand potential attacks. The security of ML methods has been previously discussed from conceptual, methodical, and practical viewpoints. Typically, the security analysis of proposed learning-based techniques is carried out informally and is occasionally supported by experimental evaluation. From the practical perspective, the success of attacks against learning algorithms crucially depends on the amount of knowledge available to an attacker. Most of the previously reported successful attacks assume that the attacker has full knowledge of the learned model.

Still, it remains largely unclear what an attacker may learn about a learning-based method deployed ‘in the wild’ and how this information can be exploited. To investigate this problem, we present the results of a case study we performed on a real learning-based model. For any submitted PDF file, the model provides a probabilistic estimate of its maliciousness. Our study addresses the case when an attacker attempts to evade detection by modifying the submitted PDF file so that its malicious functionality remains intact but the probabilistic score returned by the model is decreased.

To systematically explore the attacker’s options, we define an orthogonal set of evasion strategies reflecting various degrees of available knowledge. The general idea of our evasion technique is based on insertion of dummy content into PDF files that is ignored by PDF renderers but affects the computation of features. Once we can influence a subset of features, we develop algorithms for constructing attack instances. In the experiments, we evaluate the effectiveness of our strategies against our model.

In summary, this paper makes the following contributions:

* A new document dataset with 173,036 malicious files and 28,332 benign files.
* Identification of 133 useful and comprehensive static features for detection.
* A high accuracy rate of 99.82%, with a false positive rate of less than 0.01% for the learned model.
* Prediction time for single file maintains at a millisecond level.

Development of an adversarial examples detection framework including adversarial example generation, model hardening, evasion detection and 5 effective defence techniques.

1. Related Work

Existing malicious document detection methods can be classified broadly into two categories: dynamic and static analysis. In dynamic analysis, malicious documents are executed and examined in a specially created environment in order to capture the samples, malicious behavior; in static analysis, the detection is carried out without code execution but with static scanning and examination for the header, N-gram of files, and others. In general, the advantages of static analysis are ease of deployment and good speed (but relatively low accuracy). Compared to static analysis, dynamic analysis, although suffering from low speed and intense resource consumption, exhibits the highest accuracy. Nowadays, both techniques have already had a large number of successful stories. More advanced solutions along this line usually involve the hybrids of dynamic and static detection methods (see Maiorca *et al*. [9] for details). A summary of existing methods is presented in Table1.

Table 1. Taxonomy of malicious PDF document techniques partially based on platform diversity [8] with the addition of works after 2016 as well as summaries parsers, ML, and pattern dependencies.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Category | Focus | Detection | Work | Year | External Parser? | ML? | Discrepancy? |
| Static | JavaScript | Lexical Analysis [5] | PJScan | 2011 | Y | Y | Y |
| JavaScript | Token Clustering [12] | Vatamanu *et al*. | 2012 | Y | Y | Y |
| JavaScript | API Reference Classification [7] | Lux0r | 2014 | Y | Y | Y |
| JavaScript | Shellcode and opcode sig [13] | MPScan | 2013 | N | N | N |
| Metadata | Linearized object path [11] | PDF Malware Slayer | 2012 | Y | Y | Y |
| Metadata | Hierarchical Structure [1] | Srndic *et al*. | 2013 | Y | Y | Y |
| Metadata | Content Meta-features [24] | PDFrate | 2012 | Y | Y | Y |
| Both | Many Heuristics Combined [8] | Maiorca *et al*. | 2015 | Y | Y | Y |
| Both | Many Heuristics Combined [9] | Maiorca *et al*. | 2016 | Y | Y | Y |
| Dynamic | JavaScript | Shellcode and opcode sig [15] | MDScan | 2011 | Y | N | N |
| JavaScript | Known Attack Patterns [16] | PDF Scrutinizer | 2012 | Y | N | N |
| JavaScript | Memory Access Patterns [17] | ShellOS | 2011 | Y | N | Y |
| JavaScript | Common Maldoc Behaviors [18] | Liu *et al*. | 2014 | N | N | Y |
| JavaScript | Platform Independent Tap Point Identification [20] | tap point | 2016 | N | N | Y |
| Memory | Violation of Invariants [19] | CWXDetector | 2012 | N | N | N |
|  | OS | Platform Diversity [21] | PlatPal | 2017 | Y | N | Y |

From Table 1, we can conclude that the current focus of static analysis is JavaScript or file metadata. Typical detection techniques include Shellcode and OpCode Signature based MPScan [13], and structure and content based classification [9]. However, dynamic analysis mainly focuses on extracting the JavaScript Snippet from the file and running them directly to enable malicious behavior detection. Typical work in this line includes behavior-based analysis [20] and platform diversity-based analysis [21].

We can also see from Table 1 that all but three methods use either open-source or home-grown parsers and assume their capability. However, Carmony *et al*. [20] shows that these parsers are typically incomplete and make oversimplified assumptions about where JavaScript can be embedded. This leads to one of the most important questions in this research: Is the external parser robust enough? This is because the design and implementation of these kinds of external parsers are usually simple without being designed to be secured; In this case, only little effort is needed for the successful evasion of malicious malware. This kind of attack is called ‘parser-confusion attacks’ by Carmony *et al*. [20].

An important conclusion can be drawn from Table 1: Machine learning (ML), in general, is fit for static rather than dynamic analysis. We are not aware of a paper on dynamic analysis that considers ML, while ML has been the ‘default standard’ in nearly all static analysis papers. Typical ML work here includes PDFRATE [24] and PDF Malware Slayer [11]. Nearly all of the aforementioned works claim that their classifiers can attain high accuracy in resource-intensive environments, but seldom mention the security of their deployed ML models, much less the need for a comprehensive study of adversarial ML. This raises serious doubts about the effectiveness of classifiers based on superficial features in the presence of adversaries. An attack exploiting this vulnerability is mentioned in Xu *et al*. [14], one that can automatically produce evasive maldoc variants. Here, for each iteration and for every sample, operations like addition, deletion, and replace with respect to the PDF structure tree is performed via genetic programming like operations. During the entire process, the malicious behavior of the sample should remain the same, but the ability to confuse and evade the classifier is stronger at each iteration. This kind of attack is called a ‘classifier evasion attack’ by Xu et al. [14].

An implicit assumption is that behavioral discrepancies exist between benign and malicious documents and such discrepancies can be clearly observed and captured. Since the document must follow a public format specification, commonalities (structural or behavioral) are expected in benign documents. If a document deviates largely from the specification or the common patterns of benign samples, it is more likely to be a malicious document. In other words, a hyper-plane should always be found and posited in a high-dimensional feature space to clearly separate the malicious and benign samples. However, this assumption does not hold if one can answer the following research questions:

* Can we evade the classifier by adding, deleting, or replacing content of the malicious PDF files while keeping the malicious behaviors?
* Can we evade the classifier by gradually adding malicious content to the benign PDF?

The work by Srndic et al. [4], started with malicious files, answered the first question, which refers to as ‘mimicry attack’. Maiorca et al. [10], started with benign files, answered the second question, which refers to as ‘reverse mimicry attack’. Both works show how a malicious document can systematically evade the classifier.

In summary, for attacking an external parser, the current technique is called ‘parser-confusion attack’. For attacking ML models, the current technique is called ‘automatic classifier evasion attack’. For the assumed detectable discrepancy, the existing attack is called ‘Mimicry and Reverse Mimicry’.

There is no doubt that the attacks mentioned above attacks have raised significant challenges to the security of ML models, and potentially extends to the entire detection framework. Thus, in this paper, we propose not only some ML techniques that can achieve high accuracy, but also a truly secure, robust model under our effective defense strategies.

1. Design of ML Maldoc Classifier

In this session, we focus on designing binary classifiers, a kind of learning systems, which classify new data into two predefined categories. Classifiers usually make predictions by computing some numeric or probabilistic score and comparing it with a fixed threshold. We focus on the following aspects of classifier design:

1. General Machine Learning Classification Framework
2. Dataset
3. Feature Engineering
4. Classification Alg. Selection
5. Model Evolution
   1. General Machine Learning Classification Framework

The machine learning Classification framework we use are depicted in Figure 1. Our goal is to train a robust model for maldoc detection. Firstly, we need to collect a significant amount of malicious and benign documents during the data collection phase. Secondly, we manually design and extract hundreds of representative features from each document during our feature engineering phase, a process from transforming the documents from raw to feature vectors. Finally, we have trained the ML model so that the model can fit the underlying training data distribution well. The training phase usually performs offline due to heavy time and space footprint while the prediction phase is wrapped as online service. At this point, our model is ready for serving. When a new sample is presented to the model, it can return a confidence score to predict whether the sample is malicious or not. Figure 1 provides a good description of basic ML classification framework.

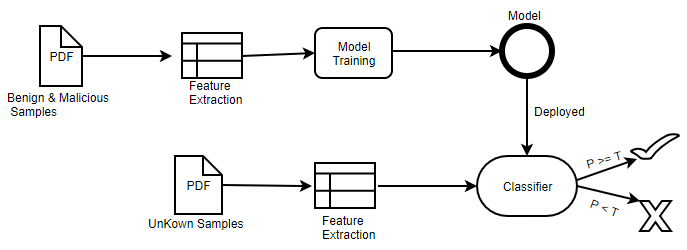


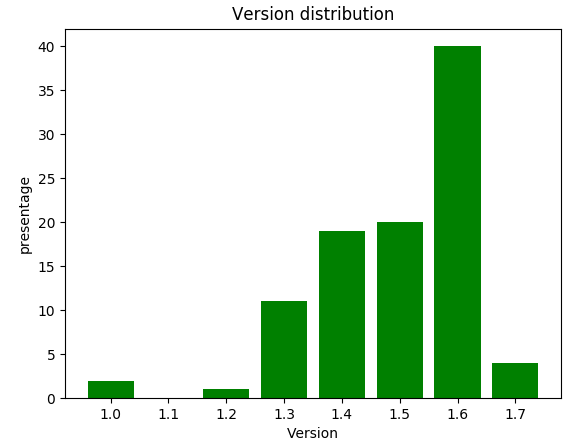
Figure 1. Basic machine learning classification framework.

* 1. Dataset

Our dataset, a total of 201,368 PDF samples, can be divided into two classes: 28,332 benign and 173,036 malicious samples.

Among those, 156,035 malicious samples are downloaded from VirusShare, 9,000 malicious samples are from the Contagio dataset, the rest are obtained from two popular search engines. Besides, we obtained the open source dataset, in the present of feature vectors, collected originally for PDFRate [4] evaluation. This dataset contains 20,000 balanced samples, with 5,000 benign and 5,000 malicious samples from Contagio dataset, and 5,000 benign samples obtained from Google as well as 5,000 malicious samples from VirusTotal. Further, we randomly select around 2,000 malicious documents from our dataset to generate around 10,000 adversarial samples targeting deployed ML model for ‘Model Evasion Attack’ in Session 4.

From Figure 2, we can see that 40% of the samples are compatible with PDF version 1.6 and the rest are mostly distributed from version 1.3 – 1.5. Lower PDF version usually means higher possibility of being exploited. Thus we suggest end users update their PDF Viewer to the latest compatible version in order to patch known vulnerabilities.



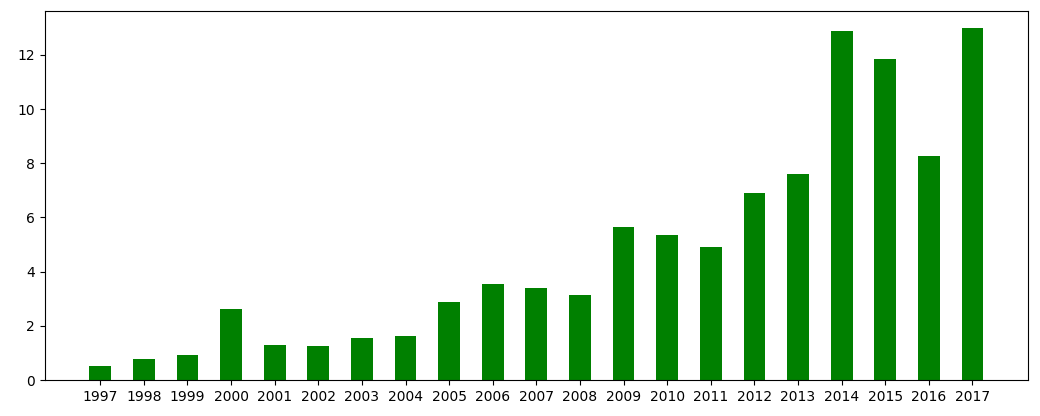
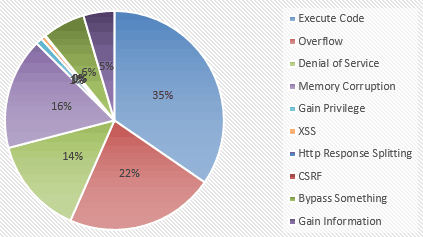
 Figure 2. Dataset divided by PDF Version

Figure 3.Dataset divided by Exploit Type Diversity

Figure 4. Dataset divided by Time

From Figure 3, we can see that 35% of the samples are belonging to the type Execute Code, followed by 22% being Overflow, 16% being Memory Corruption and 14% being Denial of Service.

From Figure 4, we plot the whole dataset by time. As shown, we collect samples till the year of 2017, followed by roughly a decreasing collection size as the year get older. Our dataset reflects the latest and complete malicious behavior changes and technology used by the malware authors.

One of the most important and interesting research problem is to evaluate the prediction capability of the ML model, specifically:

* Trained by the data from earlier versions from 1.0 to 1.5, what is the prediction capability against the version from 1.6 to 1.7?
* Trained by the data from major exploitation type such as Execute Code and Overflow, what is the prediction capability against the other exploitation types?
* Trained by the data till the year of 2010, what is the prediction capability against the most recent years?

Later, we can partially see that our ML technique works effectively against a large and diverse dataset, especially for the unseen examples, which therefore proves the generality and prediction capability of the technique we propose.

* 1. Feature Engineering

Because of the limited space of this paper, we intentionally ignore the introduction of the file format of PDF files and refer readers to the latest official open standard for PDF files. In general, PDF file format is diverse and complex. Thanks to the flexibility of the document open standard, we can even run programs in PDFs like the windows executables. Here, we ONLY focus on parsing the structure, content and metadata of the files and select features manually by our security experts. Previous research in feature engineering suggest that the combination of structure and meta based features performs well enough in terms of model prediction accuracy. We do NOT directly target the JavaScript code snippet because JavaScript code usually involves heavy dynamic analysis, facing with encrypted or obfuscated code and plus significant runtime overhead. More, we can easily infer the maliciousness of files by some basic statistics embedded in the file structure and metadata.

As shown in Table 2, we apply some basic statistics in calculating the average value of each feature and find out some representative features which the average values between benign and malicious samples are different:

Features such as count\_font and count\_box: There are several objects like font, box contained in the benign samples for content description. However, malicious files do not aim at describing information, instead they run malicious code embedded in the file to launch the attack;

Features such as count\_page\_obj and count\_obj: Generally, obj in benign files are many more than those in malicious files. When calculating the number of obj in the same page, that the malicious file is twice as many as that in a benign file. Thus, if the number of obj in the same page increases sharply, the file is likely to be malicious;

Features such as count\_endobj and count\_endstream: In benign files, the endobj refers to the end of an object. Yet a maldoc seldom contains endobj and endstream, for which it aims at confusing the parser to make it fail to obtain the whole object when parsing the malicious file, or fail to parse the malicious documents which can then evade detection successfully later;

Features such as count\_js: The main tactic of the malicious document is to embed JS code in the file to execute malicious behaviors. In this way, JS codes contained in a maldoc are generally on average, much more than those in a benign file;

Features such as count\_acroform\_obs: AcroForm is introduced in PDF Specification 1.2, which is to collect information from users via interaction. The form can display, capture and edit the data, etc. Moreover, it can conduct dynamic interaction from the interactive and editable forms which contain characteristics like dynamic calculation, verification and so on, to the forms generated by servers and filled in by machine. With those characteristics, the form is vulnerable to obscuration and encryption by the attacker. As a typical document being exploited, the value of AcroForm in a malicious sample usually doubles than that of a benign sample;

It is also worth mentioning that the features exhibit significant interdependence. When one feature’s value is modified, many others are affected because they directly or indirectly depend on each other.

At last, we manually select around 130 features for classification, in consideration of feature representative, time efficiency and robustness etc.

Table 2. Average value comparison of features in benign and malicious files.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature | | Benign file | | Malware file | |
| count\_font | | 14.64 | | 0.55 | |
| count\_acroform\_obj | | 700 | | 1400 | |
| count\_box\_a4 | | 12001 | | 200 | |
| count\_box\_legal | | 395040 | | 0 | |
| count\_box\_letter | | 7291529 | | 866773 | |
| count\_box\_other | | 32.18 | | 1.74 | |
| count\_box\_overlap | | 1000 | | 0 | |
| count\_endobj | | 95.80 | | 9.68 | |
| count\_endstream | | 30.43 | | 3.78 | |
| count\_page\_obj | | 8001 | | 16003 | |
| count\_image\_large | | 110711 | | 400 | |
| count\_image\_med | | 465247 | | 6401 | |
| count\_image\_small | | 915892 | | 12002 | |
| count\_image\_total | | 36.56 | | 0.30 | |
| count\_image\_xlarge | | 300 | | 0 | |
| count\_image\_xsmall | | 21.64 | | 0.11 | |
| count\_js | | 0.71 | | 1.01 | |
| count\_obj | | 100.96 | | 12.01 | |
| count\_objstm | | 1.57 | | 0.15 | |

* 1. Classification Alg. Selection

For classification, we select several algorithms including Decision Tree, Random Forest and SVM for comparison. As shown in Table 3, Random forest is selected for its good efficiency - minute level training time and millisecond level prediction time, good effectiveness - with accuracy as high as 99%, as well as other advantages including excellent robustness (detail in Section 4), easy to interpret and good generalization capability.

The output of random forest is essentially an ensemble of a multitude of decision trees. That said, Random forest is an ensemble classifier applying the technique of bagging training data. Each node in a decision tree is constructed based on a randomly selected subset of features, as well as the best split at each node, which is determined by training data for that node. Finally, the classification result is determined by the votes of each tree.

Table 3. Comparison of Different machine learning Models

|  |  |  |  |
| --- | --- | --- | --- |
|  | SVM | Decision Tree | Random Forest |
| Accuracy | 75.23% | 82.41% | 99.64% |
| Training Time (the whole dataset) | 58m18s | 4s | 56s |
| Prediction Time (each sample) | 1.2ms | 0.1ms | 0.1ms |

* 1. Model Evolution

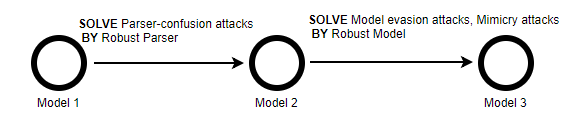


Figure 5.Model Evolution

We have 2 major updates for our model during experiment and each model provides a probabilistic estimate of the document’s maliciousness. All 3 classifiers deployed by us produce the output of their decision function, i.e., a real value in the interval [0,1] denoting the percentage of decision that have labeled the submitted file as malicious. We apply the default value of threshold (0.5) when predicting but will adjust the threshold accordingly.

**Model 1**: Use peepdf (<https://github.com/jesparza/peepdf)> as the external parser for feature extraction. These extracted features can be used for training and prediction. We extract 133 features which contain static attributes of structure such as count\_font, size and count\_startxref, content such as title\_oth and subject\_lc, metadata such as producer\_oth and producer\_len. But the limitation for peepdf is obvious: Nearly half of the PDF files can NOT be parsed correctly by the external parser. This is due to the defected file structure or intended file obfuscation technique.

**Model 2**: In this model, with the goal to conquer the major deficiency for model 1, we develop a much more robust external parser on top of the mimicus framework. By using this new external parser, nearly all the PDF files can be properly parsed. For the training of Model 2, we initially use the balanced dataset for training and testing. This balanced dataset includes 20,000 malicious and 20,000 benign examples selected randomly from the whole dataset. Besides, we extract 135 features for Model 2. The main algorithm for model is random forest. After grid search and model parameter tuning, the accuracy of Model 2 increases to 99.99%, with a false positive rate of 0.012%. We then serve our models to major commercial cloud service providers (Model-as-a-Service).

**Model 3**: The big difference between Model 2 and Model 3 is the robustness. Model 2 has the assumption that during model serving, a benign working environment is provided. But this assumption is largely not held in most situations in the context of malware detection. In Model 3, we assume that adversaries are present and adversarial examples will be submitted to launch the model evasion attacks. We will discuss this kind of attack in detail in Session 4 and propose 5 effective defense strategies we have experimented on.

1. Adversarial Machine Learning

Adversarial machine learning is a research field that lies at the intersection of machine learning and computer security [28]. It aims to enable the safe adoption of machine learning techniques in adversarial settings like spam filtering, malware detection. If the ML models are deployed online as a service, attackers can easily launch model evasion attacks or try to steal the high value model in cloud for good profit. The goal of an adversary launching the evasion attacks is to confuse the model to provide a false negative classification. Our study addresses the case when an attacker attempts to evade detection by modifying the submitted PDF files so that its malicious functionality remains intact but the probabilistic score returned by model is significantly decreased. We focus on evasion attacks in this session and leave the model stealing attacks as future work.

From an adversarial viewpoint, the more information about a learning based system is available, the higher possibility that the model can be successfully fooled. Such kind of attacks has been studied thoroughly assuming attackers have full knowledge of the deployed classifier. In practice, this assumption is rarely held, especially for the online deployed system. The knowledge about the deployed classifier can be obtained from all kinds of sources. Still, it remains largely unclear what an attacker may learn about a learning-based method deployed ‘in the wild’ and how this information can be exploited. We thus use a real, deployed ML model as a testing case, to verify the effectiveness of evasion attacks and later propose some effective model defense strategies.

We have built a framework for practical model evasion strategies as seen in Figure 6, and adapt several adversarial example algorithms for different practical application scenarios.We select particular malicious samples which are predicted as 1 by the local classifier. Later, we take advantage of different attack algorithms corresponding to different scenarios to change the feature space of the samples, making it seem to be more benign. Then prediction is made by the local trained model: with the prediction result of 1, it means the samples fail to evade; 0, means the adversarial samples successfully evade the classifier. In this way, those evaded samples can be used to test the classifier deployed on cloud. Based on the two different prediction results, that classifier will block the identified malicious samples, or let the unidentified samples evade that will finally cause actual damage to the system.

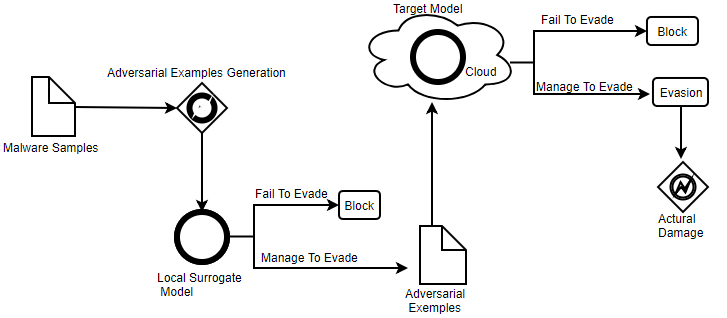


Figure 6. Model Evasion Framework

Our experiment results reveal that the detection accuracy of model 2 declines sharply from 99% to 40% even if it is exposed to simple attacks. We show that 5 potential techniques can strengthen the robustness of model against adversarial data manipulation. The rest of the session includes the following sub-sessions:

* Adversarial Examples Generation
* Case Study: How does the adversarial example work
* Potential Effective Defense Strategies
* Model Robustness
  1. Adversarial Examples Generation

In this session, we are going to discuss adversarial ML in certain scenarios. To be specific, it is supposed that an attacker has obtained some information of a targeted model, such as the features extracted, the algorithm applied and the training set. It is generally believed that as attackers know more about the model, the adversarial examples generated can perform stronger model evasion attacks. In this session, we refer to the technique proposed by Nedim Smdic [4] in order to conduct evasion attacks against the learning based model. The general idea of our evasion technique is based on insertion of dummy content into PDF files which is ignored by PDF renderers but affected the computation of features used in Model 2. Once we can influence a subset of model’s features, we develop algorithms for adversarial examples generation.

To systematically explore the attacker’s options, we define an orthogonal set of evasion strategies reflecting various degrees of available knowledge. The letters F, T or C, correspond to the feature set, training dataset and classifier algorithm respectively and are present in the name of a scenario respectively. We describe some high-level ideas for staging evasion attacks in the following 4 attack scenarios:

* Attack Scenario F

In scenario F, only the feature set is available to the adversary. The adversary might be aware of some or all features, mistakenly considering obsolete features as being used. He might also be able to read a subset or all features, or modify some or all features to a varying degree.

* Attack Scenario FT

This scenario enables the adversary to take advantage of the knowledge of the targeted classifier’s training dataset, in addition to the known features. Knowledge of training data enables the attacker to perform the entire attack offline before submitting the variants.

* Attack Scenario FC

In scenario FC, the adversary knows the feature set and some details about the classifier, such as its type, parameters or the specific implementation.

* Attack Scenario FTC

The adversary has the best chance of evading the targeted classifier if he knows the details of all three classifier components. In that case, he can fully reproduce the online classifier in an offline setting, submitting the attack results only when a sufficiently good evading sample has been found. An offline mimicry attack or an offline classifier-specific attack that defeat the offline classifier have a strong probability of defeating the online one as well.

* 1. Case Study: How does the adversarial example work

A practical way to interpret attack is to observe concrete changes in feature values produced by the attacks. Although it does not scale to cases with more features and files, this kind of investigation provides deep insight into the example at hand.

This session presents a concrete example of adversarial example. We select a file with CVE ID ‘CVE-2013-0641’ as an example. This file has the vulnerability that hackers can execute any code remotely. We apply the methodologies in the above four scenarios to vary selected samples and then check against the VirusTotal. The original sample is scanned by around 60 detection engines, within which 33 engines predict the sample as malicious. While ONLY 22 engines predict the sample variant as malicious, which demonstrates strong model evasion capability of selected adversarial example.

Table 4. Detection Accuracy of VirusTotal

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| File Hash | Origin | F | FC | FT | FTC |
| 00ba5c43b1cec186c634c24ac21982d3 cve-2013-0641 | 33/61 | 22/60 | 23/60 | 22/60 | 22/60 |

Since most PDF detection engines are based on structure and content, once we modify them, such as adding some objects from benign samples or modifying the file size, the malicious file will have a higher chance to evade the classifiers. We compare the feature space of the file BEFORE and AFTER the variation. As shown in Table 5, this ‘benignization’ is evident in the selected example. By comparing the BEFORE and AFTER columns, we see that the variation for this file includes adding an author (author\* features), setting the creation (createdate\_ts) and modification (moddate\_ts) data into recent past etc. Most changes are toward the benign class. After variation, the sample remains the malicious behavior with 10 more models being evaded.

This kind of mimicry attack is well known in security research. Its idea is to transform a malicious sample in such way that it mimics a chosen benign sample as much as possible, making the resulting mimicry sample harder to detect. This attack alg. is simple to implement, and can be applied to any classification algorithm. Besides, it does not necessarily depend on a specific learned classifier model. Thus, it is suitable for evaluation in every attack and model evasion scenario.

The results for the whole dataset reveal that even with the smallest amount of available information, say an ability to freely modify two thirds of the feature values, our attacks reduce the accuracy of the model from 99% to the most 2.92% (the FC Scenario).

Table 5 reveals the changes of feature values for a subset of features in an attack. The BEFORE column shows the feature values extracted from a malicious candidate file, the AFTER column shows how the adversarial example generation alg. transformed these values into a new data point in feature space.

Table 5. Comparison of BEFORE and AFTER feature values

|  |  |  |
| --- | --- | --- |
| Feature | BEFORE | AFTER |
| author\_lc | 0 | 6 |
| author\_len | 0 | 14 |
| author\_uc | 0 | 6 |
| count\_javascript | 1 | 6 |
| createdate\_ts | -1 | 650616173 |
| createdate\_tz | -1 | 10020 |
| moddate\_ts | -1 | 482083775 |
| keywords\_lc | 0 | 4 |
| keywords\_len | 0 | 7 |
| producer\_lc | 0 | 8 |
| producer\_len | 0 | 19 |
| version | 4 | 7 |

* 1. Potential Effective Defense Technique

One of the most important components in our framework is the adversarial examples generation algorithm. The goal is to generate PDF files whose feature vectors are likely to receive low classification scores.

We thus randomly select 2,000 highly scored malicious samples as the original seeds to generate adversarial examples. We then use these examples to attack against Model 2. The experiments in this section evaluate the effectiveness of evasion techniques presented so far. As shown in Table 6, this attack causes a great accuracy drop to model 2. In the scenario of FC, the detection accuracy of Model 2 is only 2.92%, meaning more than 97% of malicious documents are managed to evade the classifier. This motivates us the development of model 2.1, model 2.2 and model 2.3, and finally the Model 3 from the aspect of defense.

Model 2.1 is trained by increasing the size of dataset from 100K to 200K samples. Compared to Model 2, Model 2.1 has significant improvement over each of the 4 attack scenarios. But the FC and FTC case is unacceptable for practical use. The reason for the boost is that more data samples have been seen, the better generality the model will be.

Model 2.2 is trained by including a large number of adversarial examples. This technique is called adversarial training in the context of adversarial ML. Around 7000 adversarial examples have been generated and added to the training set. Compared to Model 2, Model 2.2 has large improvement over all the attack scenarios, especially for the FC and FTC case. The main reason for this is because the model now has experience of differentiating the adversarial samples in training phase, and thus has a good performance in testing as long as the training and testing sets have the same data distribution.

The improvement of Model 2.3 is its adjusted threshold. It is observed that the probability distribution for the adversarial examples have been significantly decreased and thus being mixed with the benign ones compare to their original malicious seeds. If we still apply the default original threshold (0.5), most adversarial examples will be unnoticed and thus an increase of false negative rate. By manually decreasing the detection threshold from 0.5 to 0.4, we increase the sensitivity of the model to include more suspicious files for further investigation.

The final Model 3 is the combination of Model 2.1, Model 2.2 and Model 2.3, which has been trained by a much larger dataset including a significant number of adversarial examples, and has adjusted its threshold for optimized performance. It is not surprised to see that Model 3 is superior to all the models we have evaluated so far. It is also not surprised to see that the technique adversarial training contributes the most among the 3 compared defense techniques.

Table 6. Different attack scenarios and accuracies

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Attack Scenarios | Adversarial Examples | Model 2 | Model 2.1 | Model 2.2 | Model 2.3 | Model 3 |
| F | 2000 | 71.18% | 96.71% | 89.43% | 71.03% | 98.65% |
| FC | 240\* | 2.92% | 12.50% | 81.95% | 2.85% | 88.89% |
| FT | 2000 | 84.25% | 96.76% | 98.58% | 85.89% | 98.98% |
| FTC | 2000 | 15.83% | 18.71% | 86.21% | 36.82 | 92.83% |

\*This number should be 2000 instead of 240 in an ideal experiment setting, but 240 is also good and can lead to same conclusions. We are not able to use 2000 for testing because of some technical difficulties when running the mimicus framework.

* 1. Model Robustness

In our last experiment, we have investigated the robustness of defensive mechanisms against to our evasion technique. We propose 5 defense techniques to increase the robustness of machine learning model.

If the feature sets have been exploited by attackers, we can delete the related features and retrain the machine learning models. As shown in Figure 7, the top 30 features are sorted by importance of Model 3 in ascending order. It is not surprised that features such as count\_font, count\_javascript, size, count\_obj and count\_endobj account for most of the weights. In a grey box setting of the adversarial environment, we assume the adversaries modify the 5 most important features. As an effective defense strategy, we suggest removing those exploited features and retrain the model in production as soon as possible.



Figure 7.Top 30 features

Table 7 shows the detection accuracy of the retrained Model 3 when removing top 5 features sequentially. As shown, when the classifier is trained by using all features, the accuracy rate of model is around 99.82%. If we delete up to top five features, the accuracy is almost the same. That concludes our model can confront the grey box attacks even using the rest of ‘medium important’ features for training.

Table 7. Accuracy of Model after deleting top 5 related features

|  |  |
| --- | --- |
| Features Deleted | Accuracy |
| []\* | 99.82% |
| [1] | 99.52% |
| [1,2] | 99.52% |
| [1,2,3] | 99.64% |
| [1,2,3,4] | 99.64% |
| [1,2,3,4,5] | 99.64% |

\* The top 5 related features here are: (1, count\_font), (2, count\_javascript), (3, size), (4, count\_obj) and (5, count\_endobj)

More, we assess the model robustness by the effectiveness of features. First, we sort the features by their importance, and then delete the most important features one by one. As shown in Figure 8, the curve represents the accuracy of Model 3 when features are deleted one by one. When the number of features are decreased to 100, the accuracy of the retrained model still maintains at 90%. This demonstrates:

* Despite the high weights of individual features, even they are deleted, the accuracy of model ONLY declines a small margin;
* The interaction of ‘Medium Weight’ features can make the model robust, and reduce the effect caused by the deletion of individual important features significantly;
* ‘Medium Weight’ features can help to effectively prevent the ‘Model Evasion attacks’;



Figure 8. Detection Accuracy as features are deleted one by one

Finally, in order to assess the prediction performance of Model 3, we randomly divide the dataset into two classes: 90% for training and 10% testing. 10-fold cross validation is applied. As shown in Figure 9, the area below the ROC curve is nearly 1, indicating good prediction performance.



Figure 9. ROC Curve

In summary, for the problem of model robustness, we have proposed 5 defense techniques to harden our model. They are:

• Increase the number, the diversity and the timeliness of the training dataset (as Model 2.1)

• Adversarial Training (as Model 2.2)

• Adjust the model threshold (as Model 2.3)

• Remove the features being exploited (as Model 3 in Figure 8)

• Hide classifier related info such as model type, parameters and prediction probabilities (evident enough thus no experiments are needed)

1. Application: AI Firewall

Based on the design principle of modulation, we regard the AI-based maldoc detector as an independent detection module that can be easily integrated in security products such as next-generation firewall. An important question is: How the company manages to apply AI technology seamlessly to a 30-year-old security product, featuring in static and dynamic analysis?

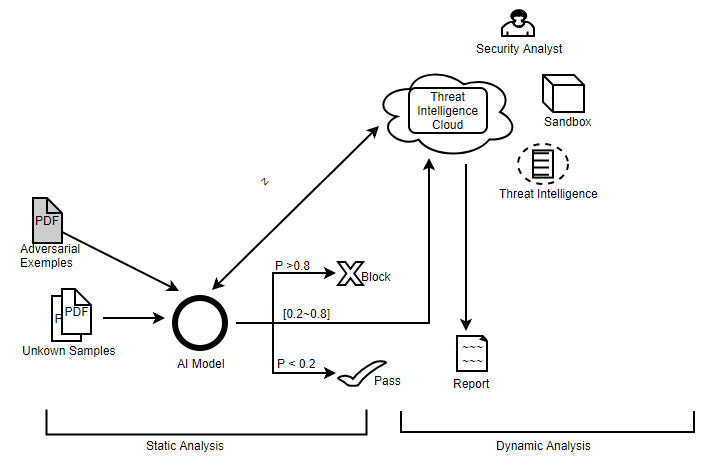


Figure 10. Overview of Static & Dynamic analysis

In the current development of network and gateway security products, the capability of performing malicious file scanning effectively and efficiently at layer 7 (the network application layer) is the international standard. The industry has strict demands on this product feature. That is, a good detection module should include (1) Millisecond latency for single file detection (2) 99% accuracy while maintaining a false positive rate of less than 0.01% and (3) The model needs to be robust when facing the adversaries.

The reason behind the low latency demand is obvious: Since a security module is ***inline*** in a working pipeline, high latency will lead to an increased packet drop rate and continually data loss. This cannot be tolerant. In the past few years, with the rapid development of malware, the former industry best practice - the pattern matching engine has gradually been used less in mainstream security applications. There are two main reasons for this: (1) In order to meet the requirement of high detection accuracy, a large number of security analysts are needed to write patterns, which is a manual process that doesn’t scale at all; and (2) Considering the fast-growing size of patterns, the time for core operation - pattern matching grows exponentially. These conditions have inspired us to discover a better engine rooted in AI.

By 2018, we managed to integrate our AI maldoc detector into a firewall, in the hope of replacing the old engine. Although both the old and new engines are based on static analysis, the improvement of shifting from the old to the new AI engine has been tremendous. First, the AI engine does not need to be updated frequently because it can detect previously unseen malware effectively for at least a year. According to our experiment results, the average update frequency for our AI engine can be as long as half a year, which is much longer if compared to the 2 week period of the old pattern matching engine. Moreover, the AI-based engine enjoys low resource consumption during execution. According to our study, during the phase of model prediction, the AI engine can only consume as much as one third of CPU and 50% of memory. The CPU portion of consumption is mainly due to the computations such as feature extraction and confidence score calculation. The memory portion of consumption is due to the fact that an AI-based model must be sited entirely in main memory when predicting.

In context of a firewall, different actions are triggered by the probability and reasons generated from the AI maldoc detection module. For instance, if the output probability is greater than a certain threshold saying 0.8, this indicates that the AI module has high confidence that this document is malicious. Thus, a blocking operation is triggered, connection is dropped, and an alert is raised requesting for further investigation. If the output probability is less than a certain threshold say 0.2, this indicates that the AI module has high confidence that this document is benign, and then the connection is allowed and monitored as normal.

The truly interesting part arises when the output probability is in the range between 0.2 and 0.8. When this happens, we by default upload the samples to the Threat Intelligence Cloud where various dynamic analysis is performed by a mixed of techniques such as Sandbox, Threat Intelligence and Security Teams. According to our preliminary experimental results, 99% of the files can be processed inline while only 1% of the files are uploaded for further investigation. Dynamic analysis from cloud plays a great complementary role for the static analysis inline. By combining the two, we can now completely provide end users with a more advanced, AI-enabled security solution. We have made our cloud service a subscription service and freely open to the research community.

1. Conclusions

In this paper, we introduce the design and implementation of a PDF file classifier based on machine learning. Our experimental results reveal that the classifier can achieve a detection accuracy greater than 99% and has a false positive rate less than 0.01%. In addition, compared with the rule based model, the time and space performance of CPU and memory improve significantly.

We than turn to the study of securing the ML models. In general, we did the following: (1) Construction of attacking scenarios in which attackers generate the adversarial examples in a way to evade the classifiers and (2) Proposing several effective defense techniques for model robustness.

We reach to an agreement that machine learning based maldoc classifier is an important research topic in context of social engineering and malware analysis. In future, we plan to strengthen our work as follows:

* Malicious documents detection based on deep learning
* Optimization of the static and dynamic analysis engines
* More file formats supported such as .doc

Acknowledgement

This work was partly supported by the School of Computer Science & Engineering, South China University of Technology and Bluedon Information Security Technologies Co, Ltd. The authors thank all the reviewers for their helpful suggestions.

References

1. Nedim ˇ Srndic and Pavel Laskov. Detection of Malicious Pdf Files Based on Hierarchical Document Structure. In 20th Network and Distributed System Security Symposium (NDSS), 2013
2. Nedim ˇ Srndic and Pavel Laskov. Mimicus: A Library for Adversarial Classifier Evasion. <https://github.com/srndic/mimicus>.
3. Nedim Šrndic and Pavel Laskov . Hidost: a static machine-learning-based detector of malicious files, EURASIP Journal on Information Security (2016) 2016
4. Nedim Srndic and Pavel Laskov. Practical Evasion of a Learning-Based Classifier: A Case Study. In Proceedings of the 35th IEEE Symposium on Security and Privacy (Oakland), San Jose, CA, May 2014
5. Pavel Laskov and Nedim Srndic. Static Detection of Malicious JavaScript-Bearing PDF Documents. In Proceedings of the Annual Computer Security Applications Conference (ACSAC), 2011
6. Davide Balzarotti, Marco Cova, Christoph Karlberger, Christopher Kruegel, Engin Kirda, and Giovann Vigna. Efficient Detection of Split Personalities in Malware. In Proceedings of the 17th Annual Network and Distributed System Security Symposium (NDSS), San Diego, CA, 2010
7. Igino Corona, Davide Maiorca, Davide Ariu, and Giorgio Giacinto. Lux0R: Detection of Malicious PDF-embedded JavaScript Code through Discriminant Analysis of API References. In Proceedings of the Artificial Intelligent and Security Workshop (AISec), 2014
8. Davide Maiorca, Davide Ariu, Igino Corona, and Giorgio Giacinto. A Structural and Content-based Approach for a Precise and Robust Detection of Malicious PDF Files. In Proceedings of the International Conference on Information Systems Security and Privacy (ICISSP), 2015.
9. Davide Maiorca, Davide Ariu, Igino Corona, and Giorgio Giacinto. An Evasion Resilient Approach to the Detection of Malicious PDF Files. In Proceedings of the International Conference on Information Systems Security and Privacy (ICISSP), 2016.
10. Davide Maiorca, IginoCorona, and Giorgio Giacinto. Looking at the Bag is not Enough to Find the Bomb: An Evasion of Structural Methods for Malicious PDF Files Detection. In Proceedings of the 8th ACM Symposium on Information, Computer and Communications Security (ASIACCS), Hangzhou, China, March 2013.
11. Davide Maiorca, Giorgio Giacinto, and Igino Corona. A Pattern Recognition System for Malicious PDF Files Detection. In Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM), 2012.
12. Cristina Vatamanu, Drago¸s Gavrilu, and R˘azvan Benchea. A Practical Approach on Clustering Malicious PDF Documents. Journal in Computer Virology, June 2012.
13. Xun Lu, Jianwei Zhuge, Ruoyu Wang, Yinzhi Cao, and Yan Chen. De-obfuscation and Detection of Malicious PDF Files with High Accuracy. In Proceedings of the 46th Hawaii International Con- ference on System Sciences (HICSS), 2013.
14. Weilin Xu, Yanjun Qi, and David Evans. Automatically Evading Classifiers: A Case Study on PDF Malware Classifiers. In Proceedings of the 2016 Annual Network and Distributed System Security Symposium (NDSS), San Diego, CA, February 2016. <http://evademl.org/>
15. Zacharias Tzermias, Giorgos Sykiotakis, Michalis Polychronakis, and Evangelos P. Markatos. Combining Static and Dynamic Analysis for the Detection of Malicious Documents. In Proceedings of the 4th European Workshop on System Security (EUROSEC), 2011.
16. Florian Schmitt, Jan Gassen, and Elmar Gerhards-Padilla. PDF Scrutinizer: Detecting JavaScript-based Attacks in PDF Documents. In Proceedings of the 10th Annual International Conference on Privacy, Security and Trust (PST), 2012.
17. Kevin Z. Snow, Srinivas Krishnan, Fabian Monrose, and Niels Provos. ShellOS: Enabling Fast Detection and Forensic Analysis of Code Injection Attacks. In Proceedings of the 20th USENIX Security Symposium (Security), San Francisco, CA, August 2011.
18. DaipingLiu, HainingWang and Angelos Stavrou. Detecting Malicious Javascript in PDF through Document Instrumentation. In Proceedings of the 44th International Conference on Dependable Systems and Networks (DSN), Atlanta, GA, 2014.
19. Carsten Willems, Felix C. Freiling, and Thorsten Holz. Using Memory Management to Detect and Extract Illegitimate Code for Malware Analysis. In Proceedings of the Annual Computer Security Applications Conference (ACSAC), 2012.
20. Curtis Carmony, Mu Zhang, Xunchao Hu, Abhishek Vasisht Bhaskar, and Heng Yin. Extract Me If You Can: Abusing PDF Parsers in Malware Detectors. In Proceedings of the 2016 Annual Network and Distributed System Security Symposium, San Diego, CA, February. NDSS 2016
21. Meng Xu and Taesoo Kim, Georgia Institute of Technology: PlatPal: Detecting Malicious Documents with Platform Diversity. 26th USENIX Security Symposium 2017
22. VirusTotal. Free Online Virus, Malware and URL Scanner. <https://www.virustotal.com/>
23. Stephan Chenette. Malicious Documents Archive for Signature Testing and Research
24. http://contagiodump.blogspot.de/2010/08/ malicious-documents-archive-for.html
25. Charles Smutz and Angelos Stavrou. Malicious PDF Detection using Metadata and Structural Features. In Proceedings of the Annual Computer Security Applications Conference (ACSAC), 2012.
26. Symantec 2017 Internet Security Threat Report <https://www.symantec.com/content/dam/symantec/docs/reports/istr-22-2017-en.pdf>
27. M.Polychronakis, K.Anagnostakis and E.Markatos. Comprehensive shellcode detection using runtime heuristics. In Annual Computer Security Applications Conference, pages 287–296, 2010.
28. Charles Smutz and Angelos Stavrou. When a Tree Falls: Using Diversity in Ensemble Classifiers to Identify Evasion in Malware Detectors. NDSS 2016
29. Adversarial Machine Learning. <https://en.wikipedia.org/wiki/Adversarial_machine_learning>

|  |  |
| --- | --- |
| Feature | Weight |
| count\_font | 0.108772 |
| count\_javascript | 0.087662 |
| size | 0.079581 |
| count\_obj | 0.069801 |
| count\_endobj | 0.059954 |
| producer\_oth | 0.05054 |
| producer\_len | 0.04837 |
| pdfid1\_num | 0.039756 |
| producer\_dot | 0.037493 |
| count\_box\_other | 0.03686 |
| count\_stream | 0.036113 |
| count\_endstream | 0.0302 |
| count\_js | 0.020743 |
| pdfid0\_len | 0.019087 |
| producer\_lc | 0.01906 |
| producer\_mismatch | 0.018668 |
| len\_obj\_max | 0.018317 |
| len\_stream\_avg | 0.017994 |
| len\_obj\_avg | 0.016931 |
| pdfid0\_num | 0.016363 |
| producer\_uc | 0.016284 |
| pdfid1\_len | 0.015949 |
| len\_stream\_max | 0.013086 |
| producer\_num | 0.011488 |
| count\_startxref | 0.009243 |
| len\_obj\_min | 0.008178 |
| count\_page | 0.008106 |
| pdfid1\_mismatch | 0.007617 |
| pdfid0\_mismatch | 0.006044 |
| createdate\_version\_ratio | 0.005681 |
| creator\_len | 0.005603 |
| moddate\_version\_ratio | 0.004428 |
| title\_oth | 0.004373 |
| creator\_lc | 0.003601 |
| box\_other\_only | 0.003494 |
| creator\_uc | 0.003382 |
| pdfid\_mismatch | 0.00239 |
| moddate\_mismatch | 0.002291 |
| count\_box\_letter | 0.002237 |
| createdate\_mismatch | 0.001802 |
| moddate\_tz | 0.0017 |
| count\_eof | 0.001628 |
| subject\_lc | 0.001521 |
| title\_num | 0.001473 |
| len\_stream\_min | 0.001459 |
| title\_len | 0.001409 |
| count\_trailer | 0.001283 |
| pdfid1\_uc | 0.001269 |
| createdate\_tz | 0.001172 |
| title\_dot | 0.00116 |
| subject\_len | 0.001011 |
| title\_uc | 0.000923 |
| version | 0.000897 |
| title\_lc | 0.000852 |
| moddate\_ts | 0.000791 |
| box\_nonother\_types | 0.000789 |
| creator\_oth | 0.000768 |
| author\_len | 0.000736 |
| count\_xref | 0.000735 |
| subject\_oth | 0.000664 |
| count\_action | 0.000659 |
| createdate\_ts | 0.00063 |
| pdfid0\_uc | 0.000599 |
| delta\_ts | 0.000594 |
| count\_acroform | 0.000591 |
| author\_uc | 0.000591 |
| image\_totalpx | 0.000526 |
| creator\_dot | 0.000494 |
| delta\_tz | 0.000457 |
| count\_objstm | 0.000434 |
| creator\_num | 0.000433 |
| author\_lc | 0.000378 |
| author\_oth | 0.000319 |
| pdfid0\_oth | 0.000309 |
| count\_stream\_diff | 0.000293 |
| count\_image\_total | 0.000258 |
| creator\_mismatch | 0.000248 |
| author\_mismatch | 0.00024 |
| title\_mismatch | 0.00023 |
| keywords\_len | 0.000219 |
| pdfid1\_oth | 0.000209 |
| author\_num | 0.000203 |
| count\_image\_large | 0.000187 |
| keywords\_lc | 0.000175 |
| pdfid0\_lc | 0.000138 |
| count\_image\_small | 0.000114 |
| keywords\_oth | 0.000109 |
| author\_dot | 8.83E-05 |
| ratio\_imagepx\_size | 8.65E-05 |
| pdfid1\_lc | 6.91E-05 |
| subject\_mismatch | 6.74E-05 |
| subject\_uc | 5.21E-05 |
| keywords\_uc | 5.03E-05 |
| keywords\_mismatch | 4.02E-05 |
| count\_javascript\_obs | 2.06E-05 |
| count\_page\_obs | 1.51E-05 |
| count\_box\_a4 | 1.24E-05 |
| image\_mismatch | 1.02E-05 |
| subject\_dot | 9.83E-06 |
| count\_image\_med | 9.75E-06 |
| count\_font\_obs | 9.20E-06 |
| count\_action\_obs | 8.90E-06 |
| company\_mismatch | 7.60E-06 |
| keywords\_num | 7.54E-06 |
| count\_js\_obs | 6.23E-06 |
| subject\_num | 5.21E-06 |
| count\_image\_xsmall | 3.92E-06 |
| count\_acroform\_obs | 0 |
| count\_box\_legal | 0 |
| count\_box\_overlap | 0 |
| count\_image\_xlarge | 0 |
| count\_objstm\_obs | 0 |
| createdate\_dot | 0 |
| keywords\_dot | 0 |
| moddate\_dot | 0 |
| pdfid0\_dot | 0 |
| pdfid1\_dot | 0 |
| pos\_acroform\_avg | 0 |
| pos\_acroform\_max | 0 |
| pos\_acroform\_min | 0 |
| pos\_box\_avg | 0 |
| pos\_box\_max | 0 |
| pos\_box\_min | 0 |
| pos\_eof\_avg | 0 |
| pos\_eof\_max | 0 |
| pos\_eof\_min | 0 |
| pos\_image\_avg | 0 |
| pos\_image\_max | 0 |
| pos\_image\_min | 0 |
| pos\_page\_avg | 0 |
| pos\_page\_max | 0 |
| pos\_page\_min | 0 |
| ratio\_size\_obj | 0 |
| ratio\_size\_page | 0 |
| ratio\_size\_stream | 0 |

**Biography**

**Yubin Yang** is currently Ph.D. candidate at the School of Computer Science & Engineering, South China University of Technology, China. He previously earned a M. Sc. Degree at the University of New South Wales, Australia in 1999 and a B.Sc. degree at Guangdong University of Technology, China in 1997. His current research interests focus on information security.

**Wei Jiang** is a Ph.D. Candidate in the Department of Computer Science and Engineering at the NYU Tandon School of Engineering in Brooklyn, NY. He received a Bachelor's degree in computer science from Guangdong University of Technology in China, and a Master's degree in computer science from University of Southern California in US.

**Fengjiao Wang** graduated from the School of Computer Science and Technology, Southwest Forestry University, for the degree of Bachelor. Her research interests include Malicious Document Detection and Machine Learning.

**Shumin Wei** received the Bachelor degree from Guangzhou University in China, and the Master degree from the Hong Kong Polytechnic University. Her current research interests include corpus-based translation and machine-aided translation.