

GVHD: THS. PHAN THẾ DUY  
NT230.O21.ANTT – BÁO CÁO CUỐI KỲ

# PAPER: APPLYING NLP TECHNIQUES TO MALWARE DETECTION IN A PRACTICAL ENVIRONMENT

GROUP 17

NGUYỄN PHƯƠNG TRINH – 21521581

NGUYỄN THỊ MINH CHÂU – 21520645

TRẦN MINH DUY - 21522010



# NỘI DUNG BÁO CÁO

1  
GIỚI THIỆU  
ĐỀ TÀI

3  
TRIỂN KHAI  
PHƯƠNG PHÁP

2  
PHƯƠNG PHÁP  
NGHIÊN CỨU

4  
SO SÁNH  
VÀ ĐÁNH GIÁ



# NỘI DUNG BÁO CÁO

1  
GIỚI THIỆU  
ĐỀ TÀI

3  
TRIỂN KHAI  
PHƯƠNG PHÁP

2  
PHƯƠNG PHÁP  
NGHIÊN CỨU

4  
SO SÁNH  
VÀ ĐÁNH GIÁ

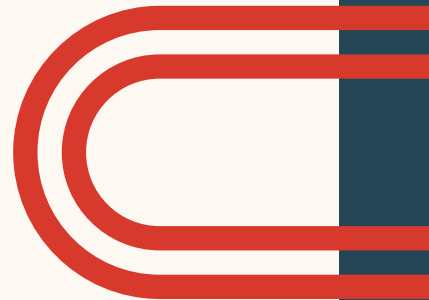
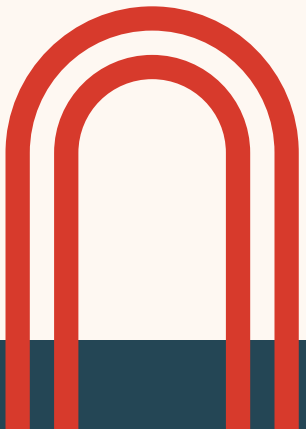
# GIỚI THIỆU ĐỀ TÀI

## NGŨ CẢNH

Các tệp thực thi (executable files) thường được sử dụng để tấn công máy tính đầu cuối và được che giấu để tránh bị phát hiện.

Phân tích động tất cả các tệp đáng ngờ từ Internet mất rất nhiều thời gian, do đó cần một phương pháp lọc nhanh chóng.

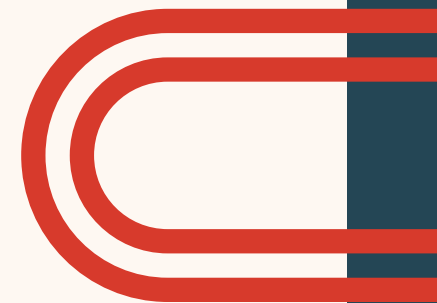
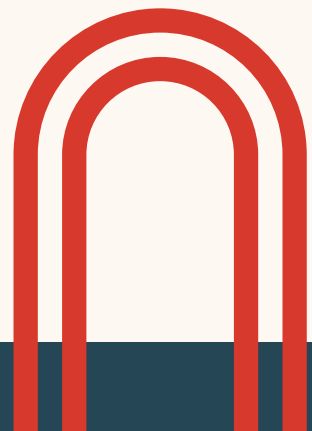
Bài báo này áp dụng kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) và chuỗi ký tự có thể in ra (printable strings) để phát hiện phần mềm độc hại.



# GIỚI THIỆU ĐỀ TÀI

## NGŨ CẢNH

Kết quả cho thấy phương pháp này hiệu quả trong việc phát hiện cả phần mềm độc hại hiện có và mới xuất hiện sau này, bao gồm cả phần mềm độc hại được đóng gói và sử dụng kỹ thuật chống gỡ lỗi, với bộ dữ liệu hơn 500.000 mẫu.



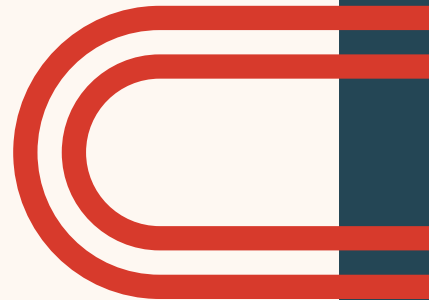
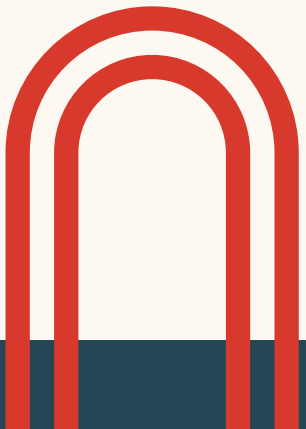
# GIỚI THIỆU ĐỀ TÀI

## GIẢI PHÁP CỦA BÀI BÁO

Với sự phát triển của các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP), sự kết hợp giữa các printable strings và các kỹ thuật NLP có thể được sử dụng như một phương pháp phát hiện malware.

*---Printable strings đề cập đến các chuỗi ký tự trong một tệp thực thi có thể được in ra hoặc hiển thị.---*

Bằng cách áp dụng các kỹ thuật NLP để phân tích và xử lý các printable strings này, có thể trích xuất thông tin và patterns có ý nghĩa, từ đó giúp xác định và phát hiện phần mềm độc hại.





# NỘI DUNG BÁO CÁO

1  
GIỚI THIỆU  
ĐỀ TÀI

3  
TRIỂN KHAI  
PHƯƠNG PHÁP

2  
PHƯƠNG PHÁP  
NGHIÊN CỨU

4  
SO SÁNH  
VÀ ĐÁNH GIÁ

# PHƯƠNG PHÁP NGHIÊN CỨU

Bài báo sử dụng một số kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) liên quan đến nghiên cứu này.

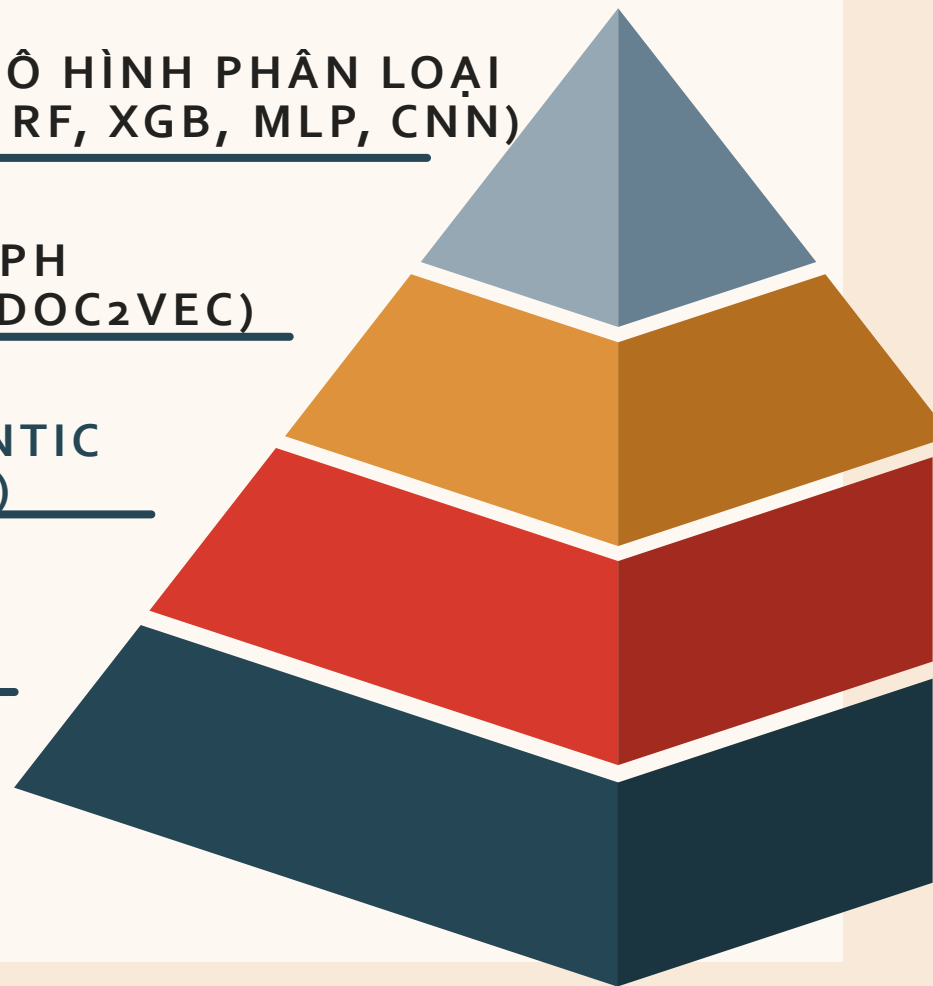
Mục tiêu chính của các kỹ thuật NLP là giúp máy tính xử lý và phân tích lượng lớn dữ liệu ngôn ngữ tự nhiên.

BAG-OF-WORDS

LATENT SEMANTIC INDEXING (LSI)

PARAGRAPH VECTOR (DOC2VEC)

CÁC MÔ HÌNH PHÂN LOẠI (SVM, RF, XGB, MLP, CNN)





# PHƯƠNG PHÁP NGHIÊN CỨU

## BAG-OF-WORDS (BoW)

Bag-of-Words (BoW) là một phương pháp cơ bản trong phân loại tài liệu, sử dụng tần suất của mỗi từ làm đặc trưng để huấn luyện bộ phân loại. Mô hình này chuyển mỗi từ trong tài liệu thành vector dựa trên tần suất của từ đó. Tài liệu được biểu diễn dưới dạng vector (ma trận tài liệu từ) để ghi lại tần suất (không theo thứ tự) của tất cả các từ khác biệt.

I love this movie! It's sweet,  
but with satirical humor. The  
dialogue is great and the  
adventure scenes are fun...  
It manages to be whimsical  
and romantic while laughing  
at the conventions of the  
fairy tale genre. I would  
recommend it to just about  
anyone. I've seen it several  
times, and I'm always happy  
to see it again whenever I  
have a friend who hasn't  
seen it yet!

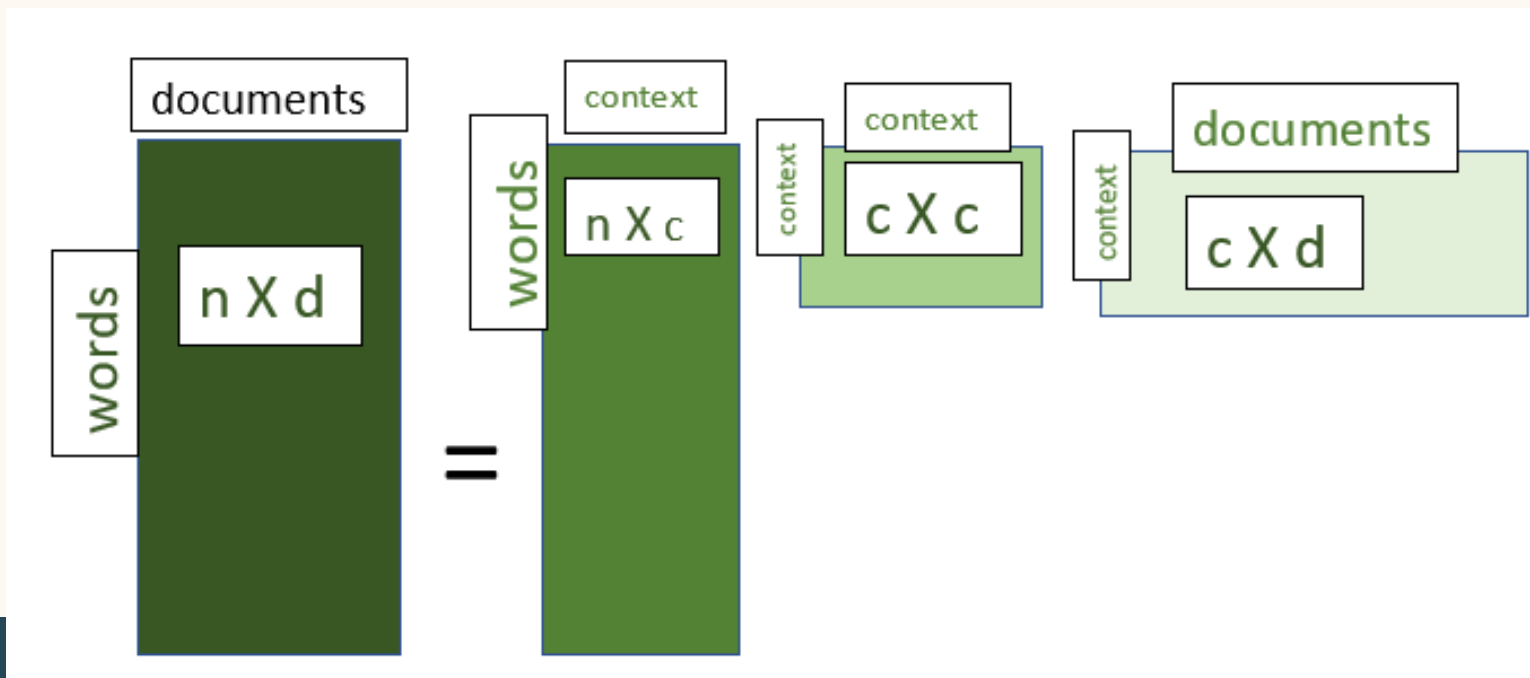


it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

# PHƯƠNG PHÁP NGHIÊN CỨU

## LATENT SEMANTIC INDEXING (LSI)

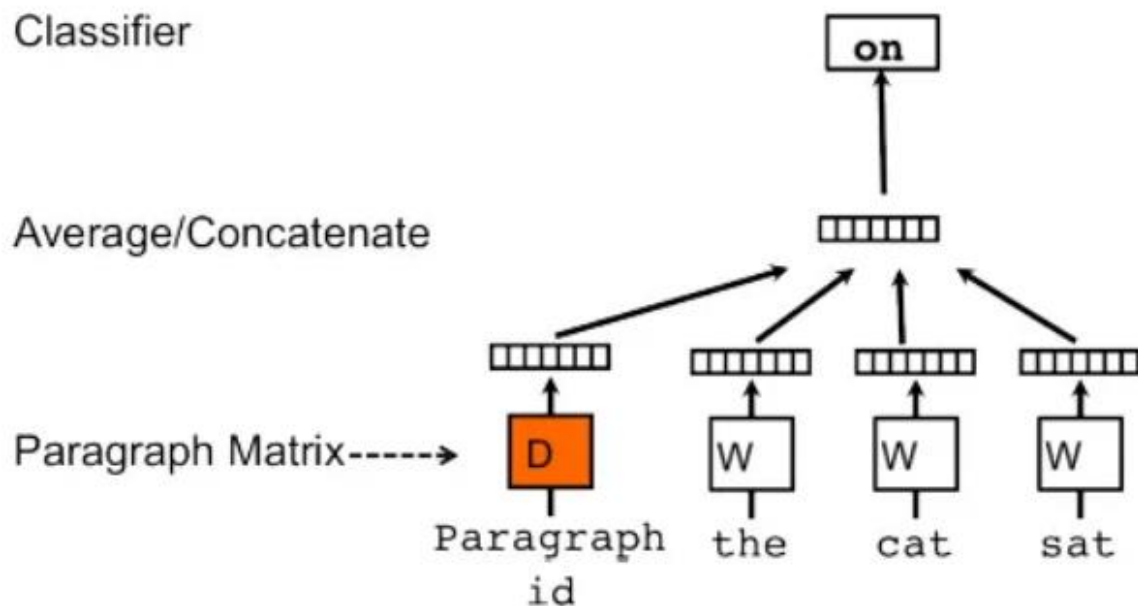
Latent Semantic Indexing (LSI) phân tích mối quan hệ giữa nhóm tài liệu và các từ trong tài liệu. Trong mô hình LSI, các vector sử dụng Bag-of-Words (BoW) được nhân với trọng số Term Frequency-Inverse Document Frequency (TF-IDF), sau đó giảm bớt không gian vector bằng phương pháp phân tích giá trị suy biến (SVD). Ma trận phân tích cho thấy mối quan hệ giữa nhóm tài liệu và các từ trong tài liệu.



# PHƯƠNG PHÁP NGHIÊN CỨU

## PARAMETER VECTOR (Doc2vec)

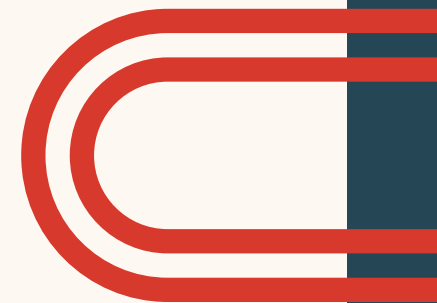
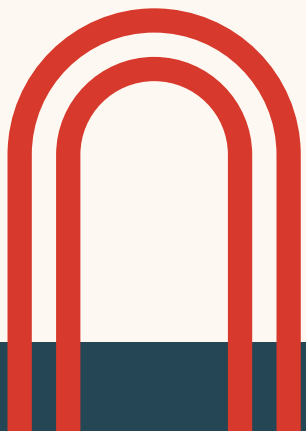
Không như BoW tập trung vào tần suất từ và không lưu giữ được ngữ cảnh, Doc2vec tạo ra vector biểu diễn ngữ cảnh toàn diện cho cả tài liệu. Phương pháp này giúp mô hình hiểu được ngữ cảnh tổng thể của tài liệu, cung cấp thông tin ngữ nghĩa phong phú hơn cho quá trình phân loại.



# PHƯƠNG PHÁP NGHIÊN CỨU

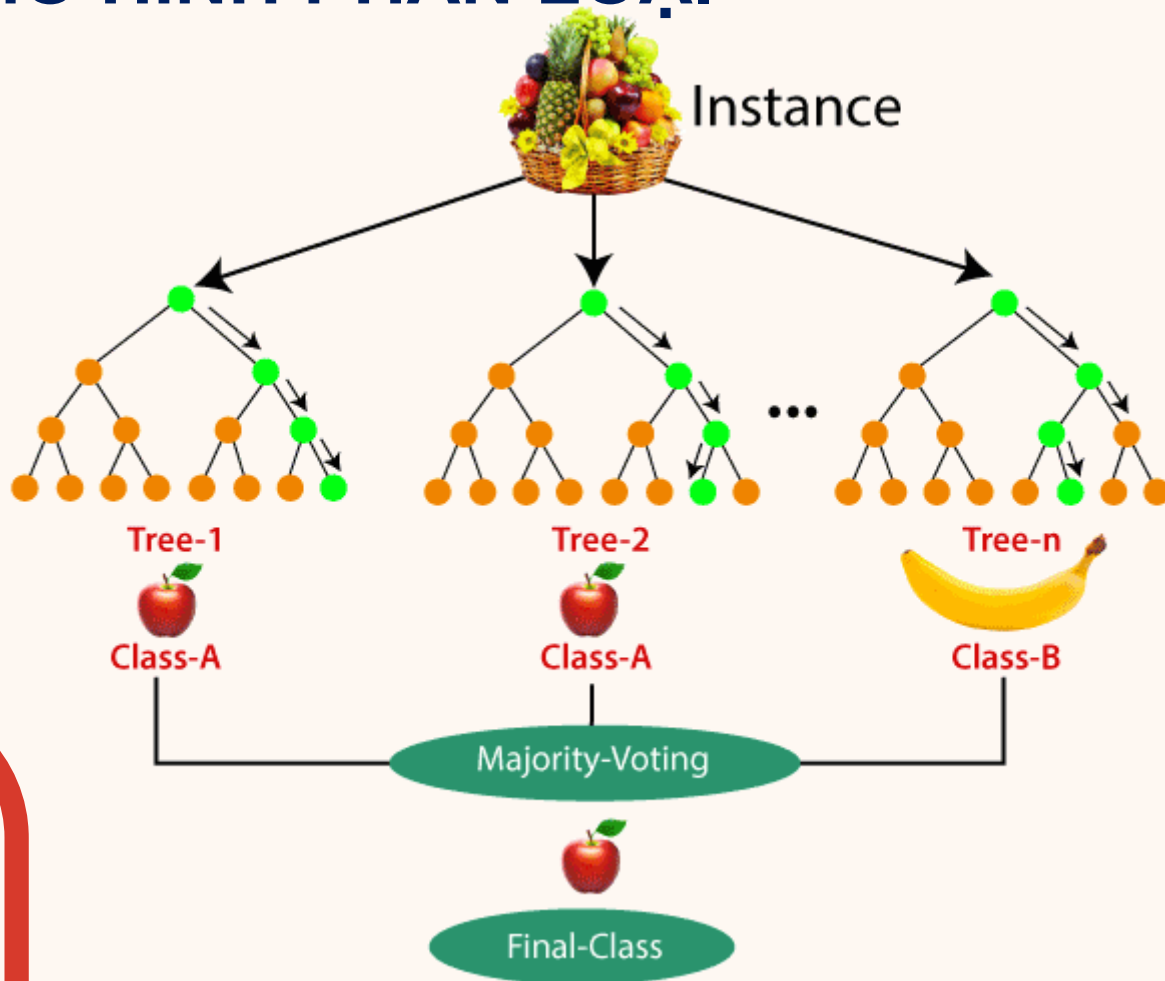
## CÁC MÔ HÌNH PHÂN LOẠI

Sau khi chuyển đổi các chuỗi ký tự thành vector đặc trưng bằng các kỹ thuật trên, các mô hình phân loại khác nhau như Support Vector Machine (SVM), Random Forest (RF), XGBoost (XGB), Multi-Layer Perceptron (MLP), và Convolutional Neural Network (CNN) được sử dụng để dự đoán nhãn của các mẫu phần mềm. Mỗi mô hình có cách tiếp cận riêng để xử lý và phân loại dữ liệu, đảm bảo độ chính xác cao trong việc phát hiện phần mềm độc hại.



# PHƯƠNG PHÁP NGHIÊN CỨU

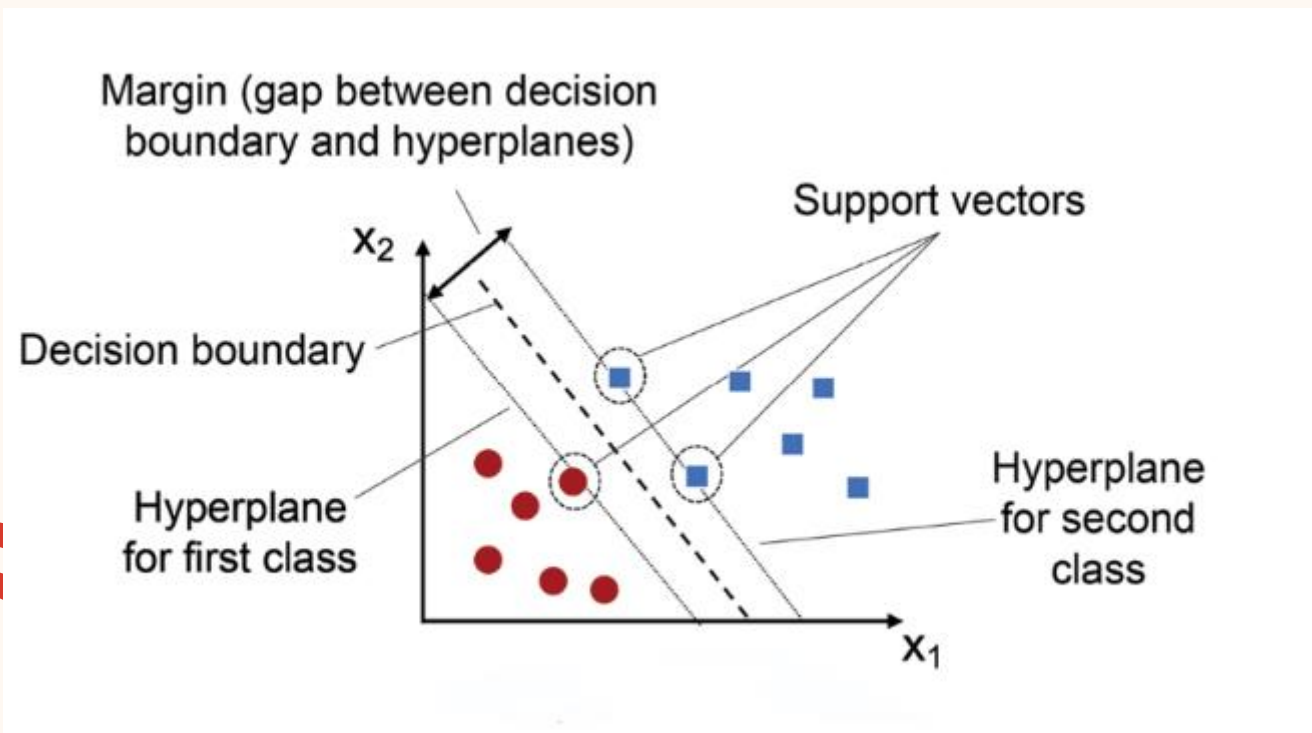
## CÁC MÔ HÌNH PHÂN LOẠI



Random Forest

# PHƯƠNG PHÁP NGHIÊN CỨU

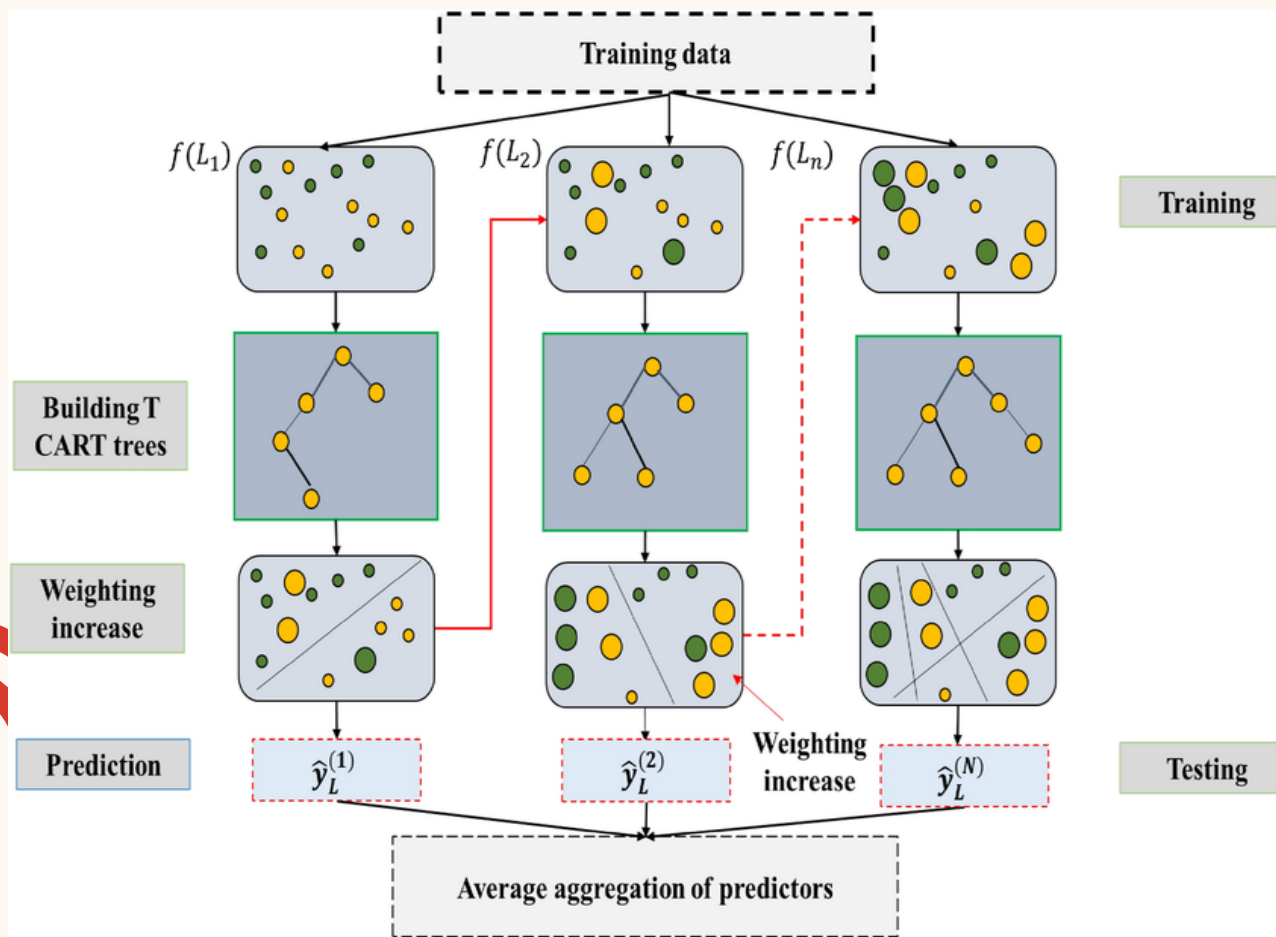
## CÁC MÔ HÌNH PHÂN LOẠI



**Support Vector  
Machine (SVM)**

# PHƯƠNG PHÁP NGHIÊN CỨU

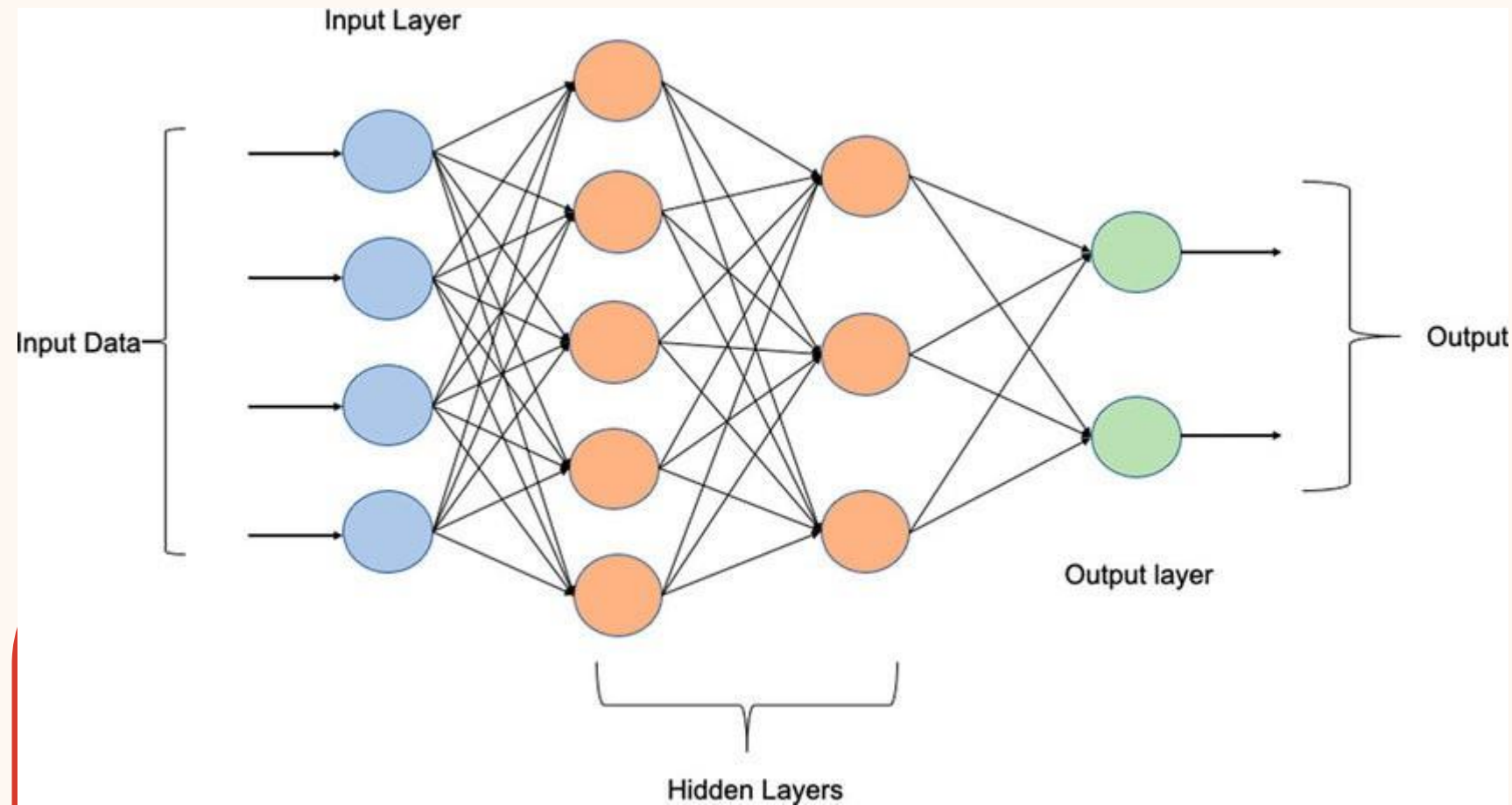
## CÁC MÔ HÌNH PHÂN LOẠI



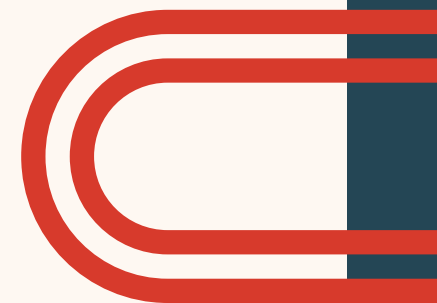
XGBoost

# PHƯƠNG PHÁP NGHIÊN CỨU

## CÁC MÔ HÌNH PHÂN LOẠI



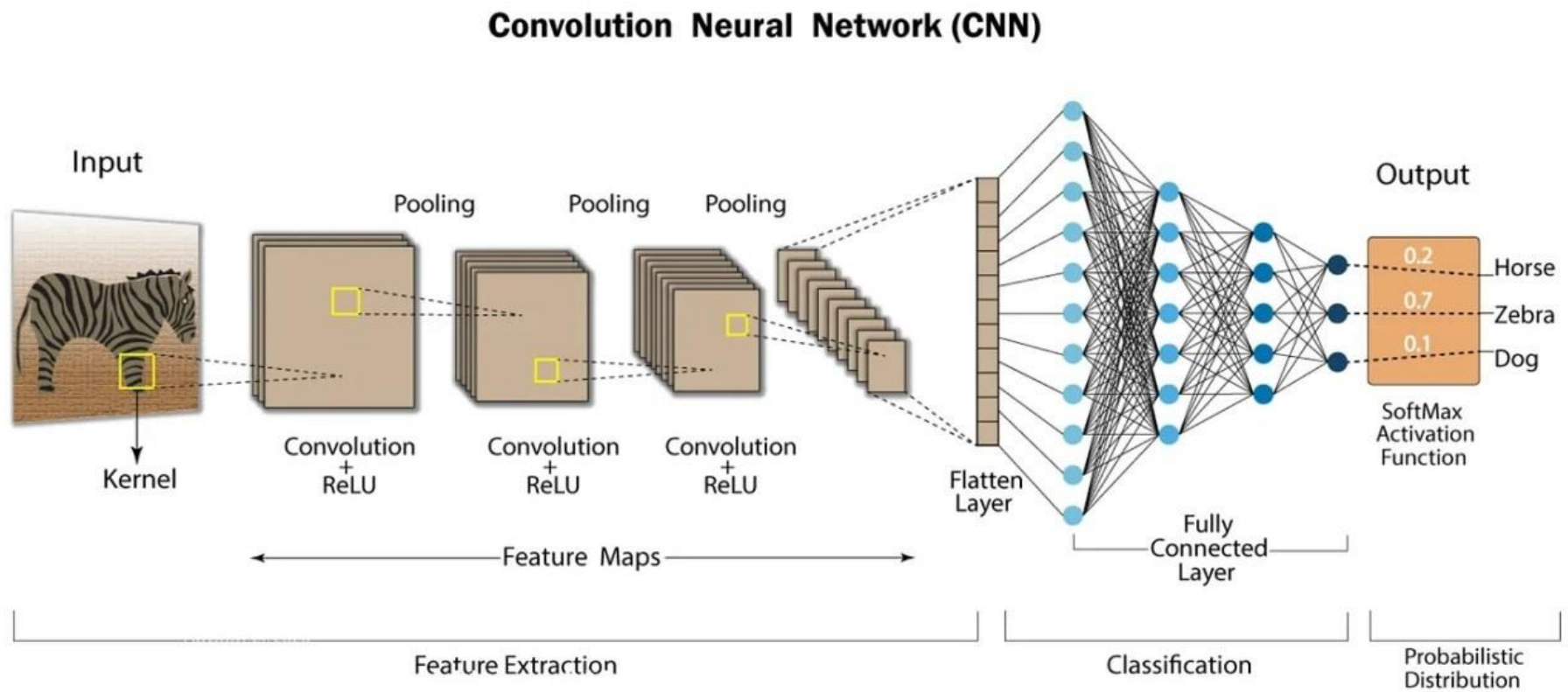
**MLP**





# PHƯƠNG PHÁP NGHIÊN CỨU

## CÁC MÔ HÌNH PHÂN LOẠI

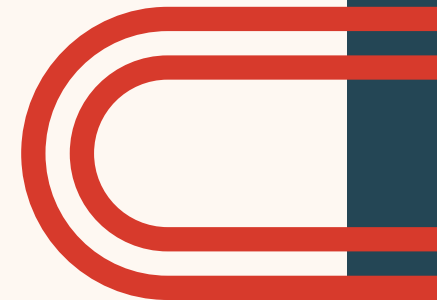
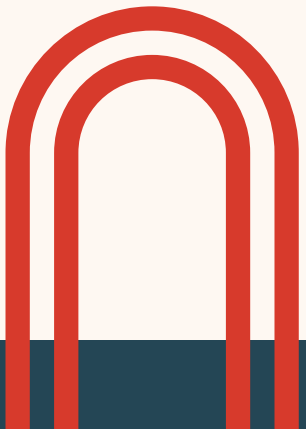
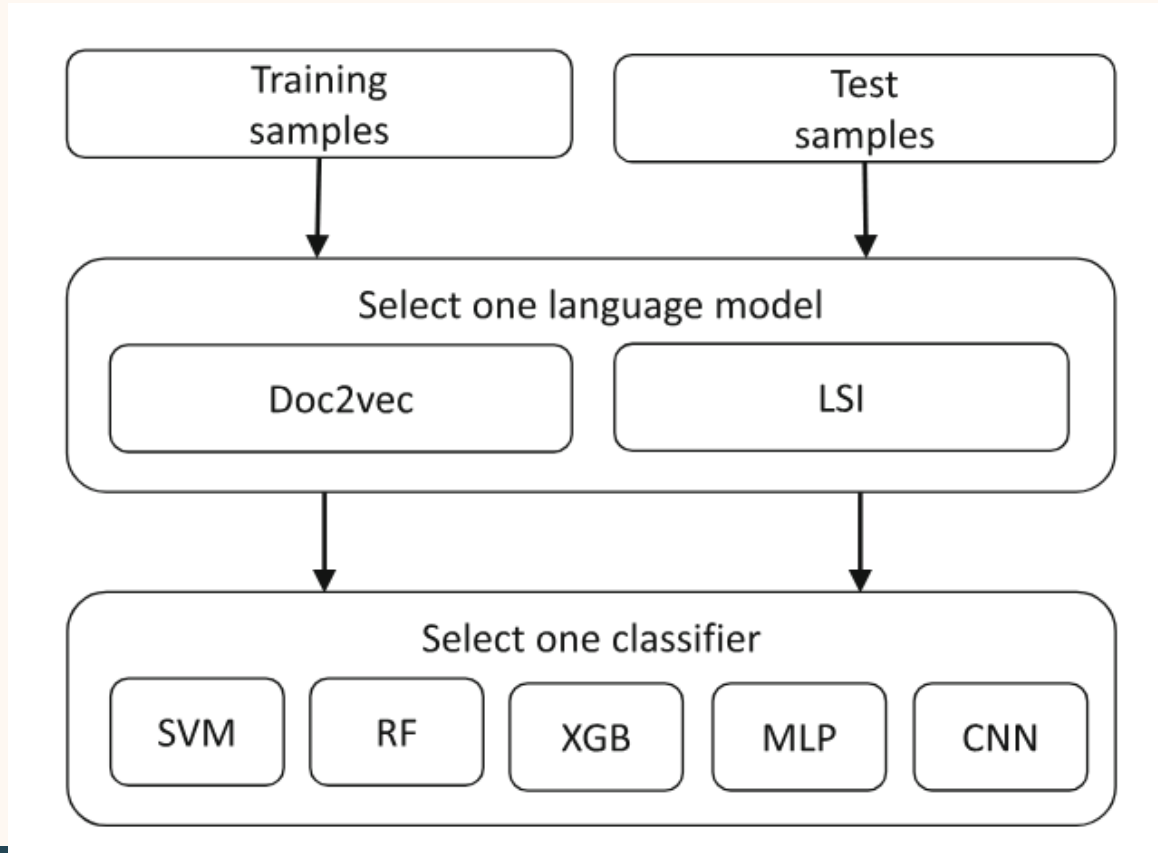


CNN

# PHƯƠNG PHÁP NGHIÊN CỨU

## CÁC MÔ HÌNH PHÂN LOẠI

Các kỹ thuật này phối hợp với nhau để tạo thành một phương pháp phát hiện phần mềm độc hại hiệu quả, từ việc trích xuất đặc trưng cho đến quá trình phân loại và dự đoán.



# PHƯƠNG PHÁP NGHIÊN CỨU

## TRAINING

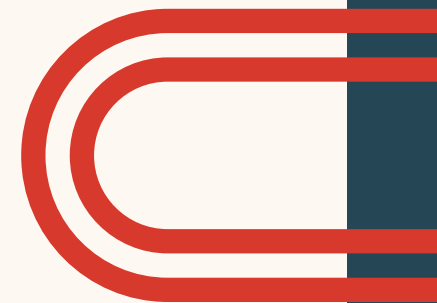
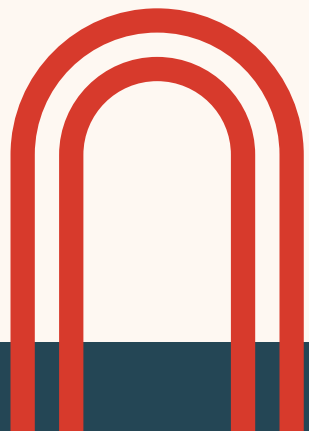
Phương pháp:

- Trích xuất tất cả các chuỗi ký tự có thể in ra (ASCII) từ các sample → tách thành các từ.
- Các từ xuất hiện thường xuyên sẽ được chọn để xây dựng mô hình huấn luyện.
- Sử dụng mô hình ngôn ngữ Doc2vec hoặc LSI để biểu diễn các đặc trưng của từ. Trong đó, Doc2vec được xây dựng từ tập hợp các từ, còn LSI được xây dựng từ kĩ thuật TF-IDF với các từ đã lấy ra.
- Các từ này sau đó được chuyển đổi thành vector từ vựng đặc trưng bằng mô hình ngôn ngữ đã chọn.
- Các vector đặc trưng sau đó được sử dụng để huấn luyện các bộ phân loại, gồm Support Vector Machine (SVM), Random Forests (RF), XGBoost (XGB), Multi-Layer Perceptron (MLP), và Convolutional Neural Networks (CNN).

# PHƯƠNG PHÁP NGHIÊN CỨU

## TESTING

Phương pháp: Sử dụng tập dataset hoàn toàn mới (như trong paper thì tác giả để model học phân loại trên FFRI 2013 – 2015 và test trên tập chứa các malware mới, chưa được sử dụng trong quá trình học). Tương tự, trích xuất lấy các vector từ vệt đặc trưng từ tập dataset mới này và đưa vào model đã học để phân loại và đánh giá



# PHƯƠNG PHÁP NGHIÊN CỨU

## MÔI TRƯỜNG THỰC NGHIỆM

- **Cấu hình máy tính:** Hệ điều hành Windows 10, CPU Intel Core i7-5820K 3.3GHz, RAM 32GB DDR4 và ổ cứng Serial ATA 3 HDD
- **Ngôn ngữ lập trình:** Python 2.7
- **Các công cụ hỗ trợ sẵn có** trong giai đoạn hiện thực phương pháp:
  - Gensim: cung cấp các mô hình LSI và Doc2vec
  - Scikit-learn: cung cấp các bộ phân loại SVM và RF
  - XGBoost: cung cấp bộ phân loại XGB
  - Chainer: dùng để hiện thực MLP và CNN
  - PEiD: công cụ phát hiện hầu hết các kỹ thuật đóng gói và chống gỡ lỗi phổ biến trong tệp PE.

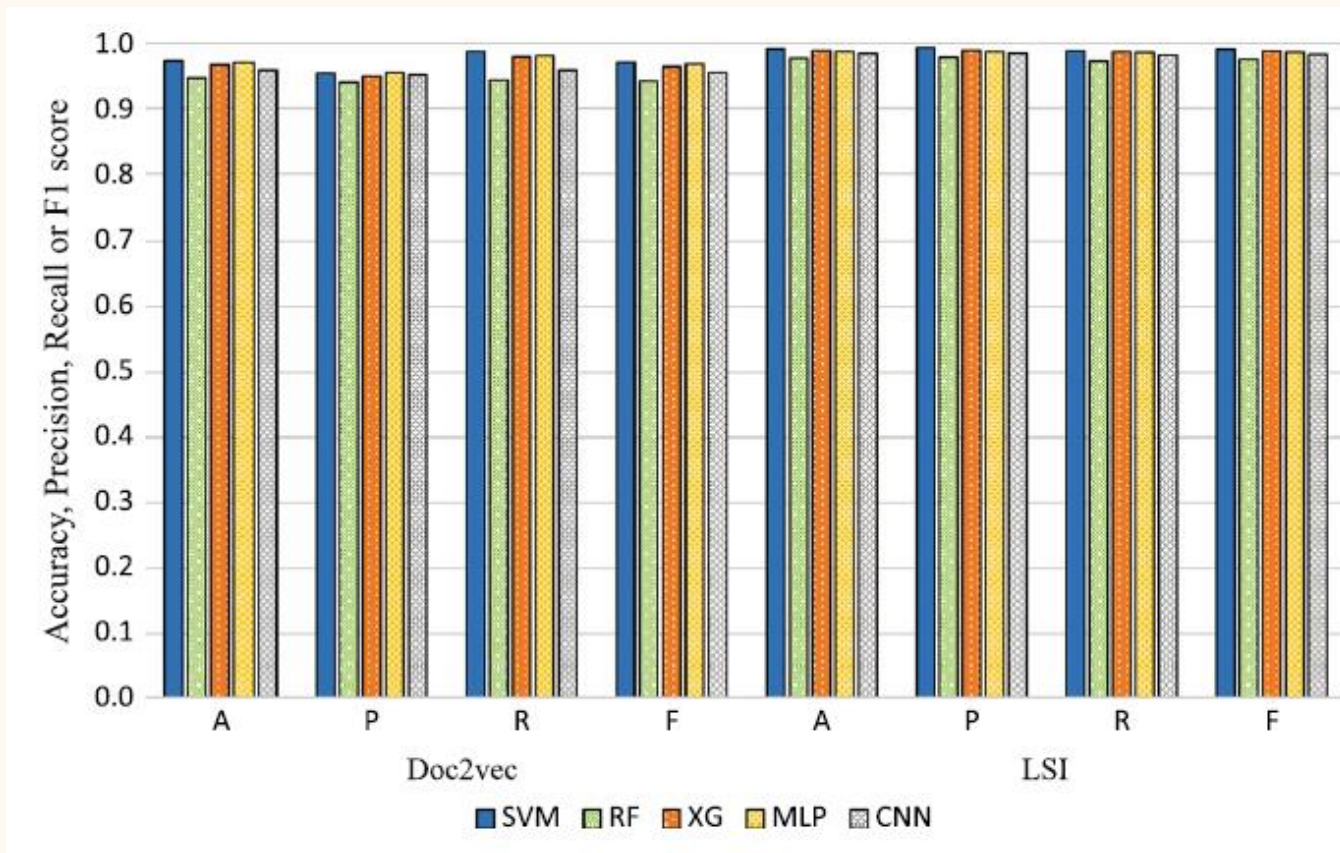
# PHƯƠNG PHÁP NGHIÊN CỨU

## MÔI TRƯỜNG THỰC NGHIỆM

- **Đối tượng nghiên cứu:** tập dữ liệu bao gồm hơn 500,000 mẫu từ các nguồn khác nhau là FFRI dataset, Hybrid Analysis. Tuy nhiên tập dữ liệu FFRI dataset nhóm em không xin được quyền truy cập từ tác giả nên đã tự thực hiện thu thập dataset theo phương pháp giống với bài báo đề ra.
- **Tiêu chí đánh giá** tính hiệu quả của phương pháp: Accuracy, Precision, Recall, F1 score, Receiver Operating Characteristics (ROC) curve, Area under the ROC Curve (AUC), thời gian đào tạo và kiểm tra

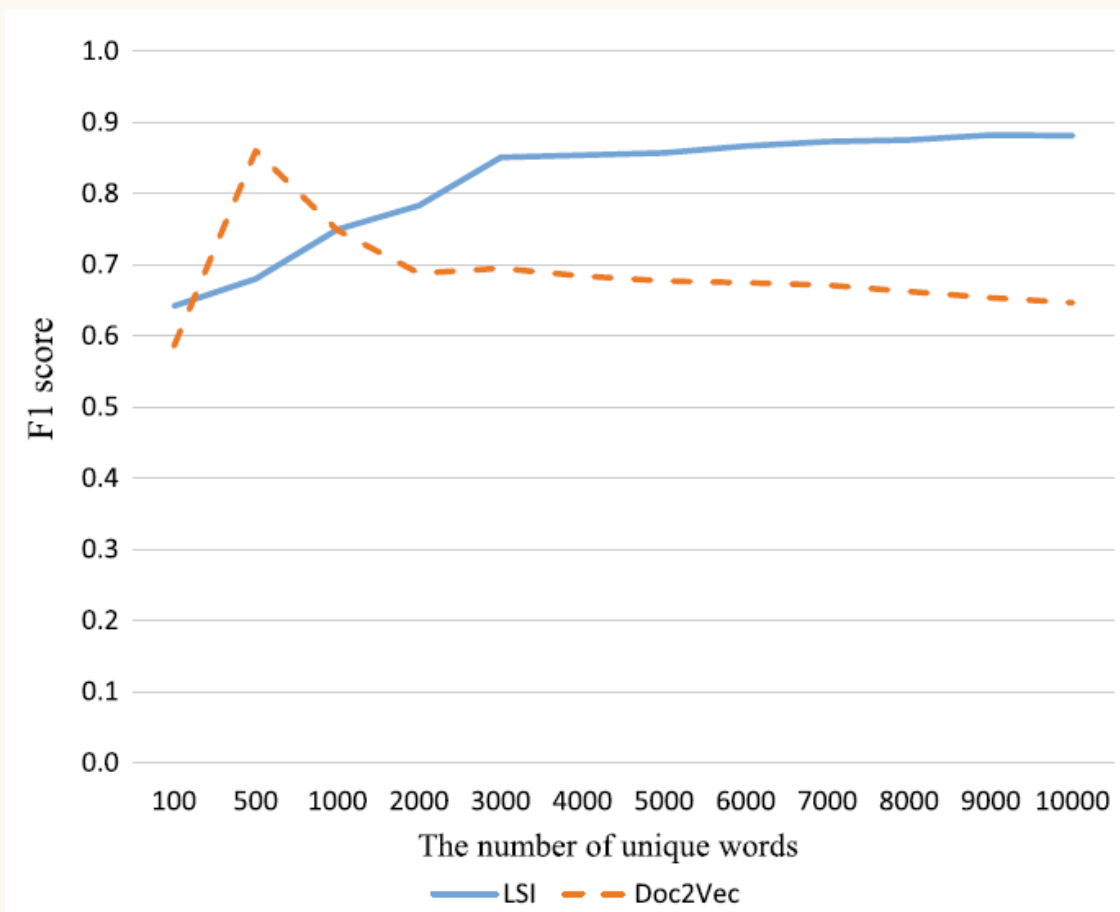
# PHƯƠNG PHÁP NGHIÊN CỨU

## KẾT QUẢ THỰC NGHIỆM



# PHƯƠNG PHÁP NGHIÊN CỨU

## KẾT QUẢ THỰC NGHIỆM





# PHƯƠNG PHÁP NGHIÊN CỨU

## KẾT QUẢ THỰC NGHIỆM

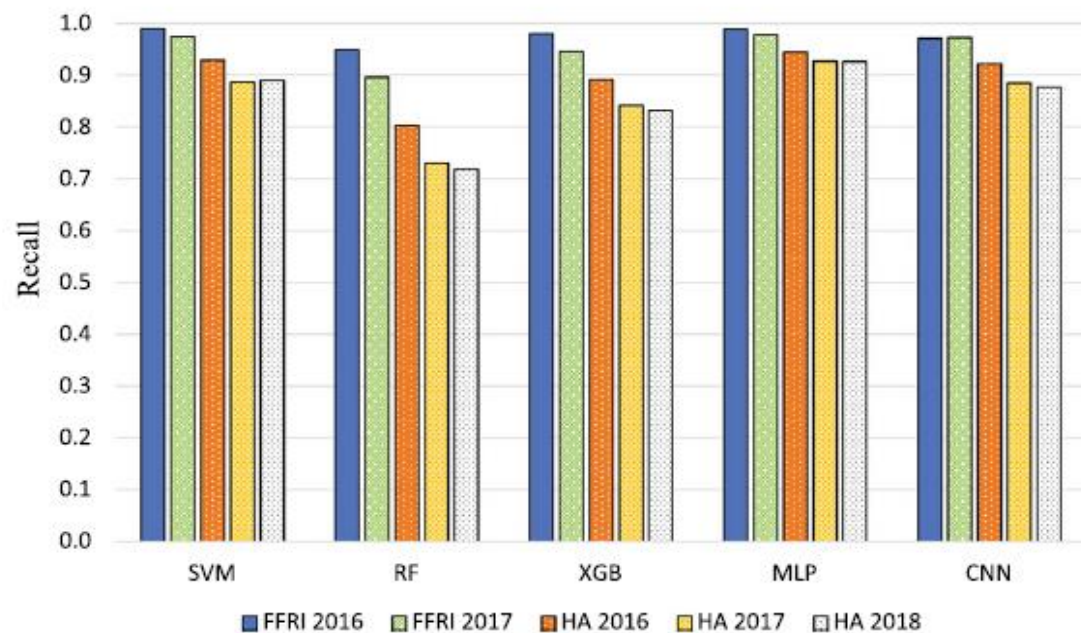


Fig. 5 The result of LSI in the time series analysis

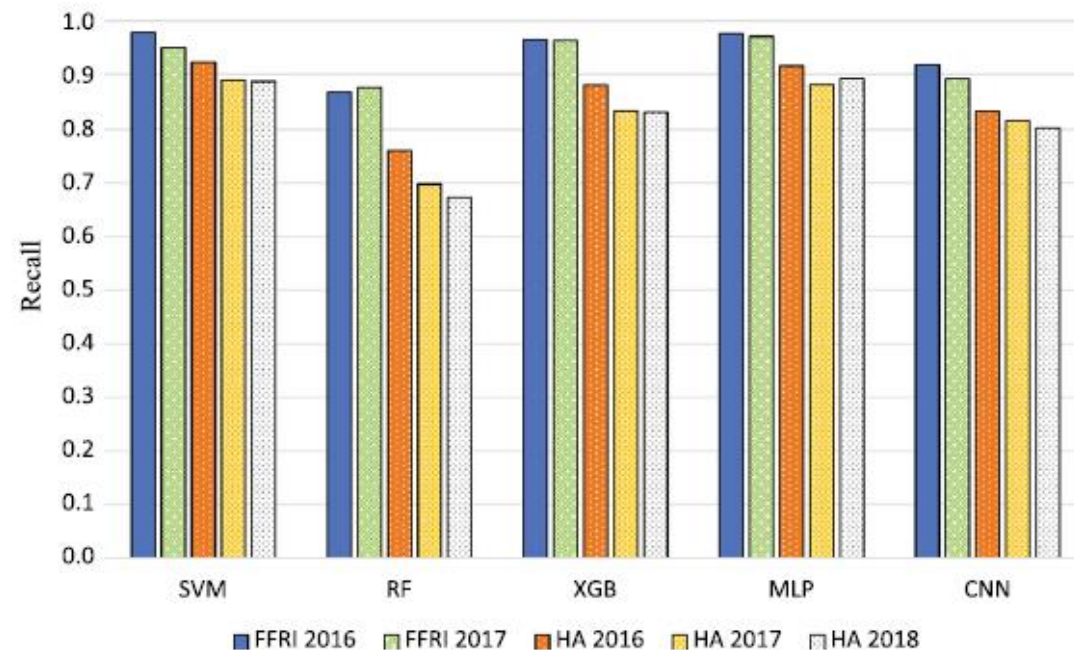


Fig. 4 The result of Doc2vec in the time series analysis



# NỘI DUNG BÁO CÁO

1  
GIỚI THIỆU  
ĐỀ TÀI

3  
THỰC NGHIỆM

2  
PHƯƠNG PHÁP  
NGHIÊN CỨU

4  
SO SÁNH  
VÀ ĐÁNH GIÁ

# THỰC NGHIỆM

## TẠO DATASET

Bài báo cung cấp script để tạo ra dataset với định dạng giống với dataset FFRI

### FFRI Dataset scripts

---

This script enables you to create datasets in the same format as the FFRI dataset.

### Requirements

---

We recommend that you use Docker for making datasets. See [Using Docker](#) for more details.

Alternatively, you can use this script by installing the following dependencies on [tested platforms](#). See [Run this script natively](#) for more details.

- Python 3.11
- [Poetry](#) 1.2+

# THỰC NGHIỆM

## TẠO DATASET

Để chạy script thì trước hết cần phải có các file sample và tạo file csv làm input với định dạng theo yêu cầu

### Make A CSV File

This script requires a CSV file which contains file information such as labels, dates, file paths. For instance,

```
path,label,date
./data/cleanware/test0.exe,0,2018/01/01
./data/malware/test1.exe,1,2018/01/02
```

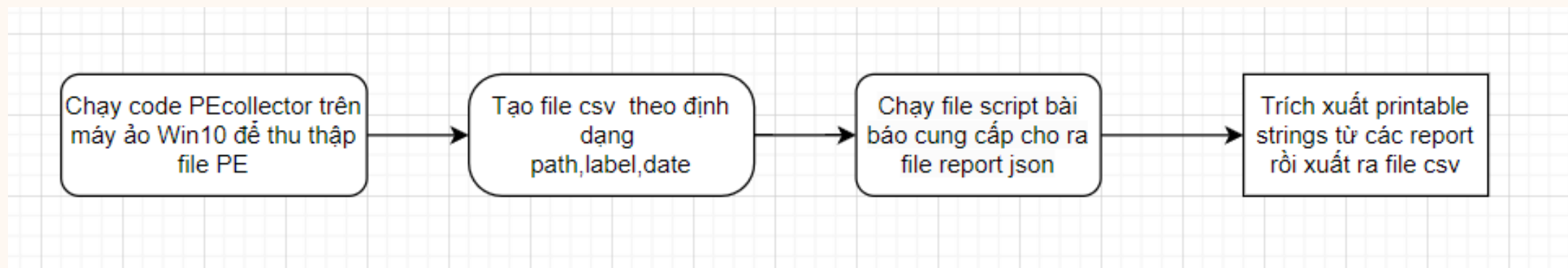


Note that file paths in a CSV file should be specified as relative paths to the container's working directory.

# THỰC NGHIỆM

## TẠO DATASET

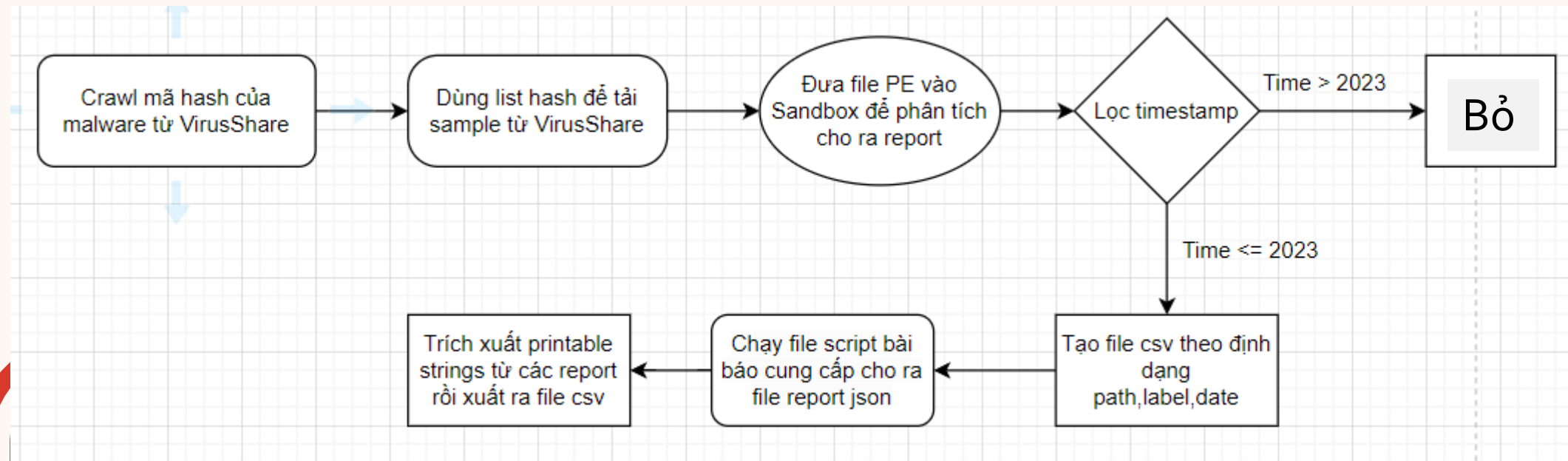
Nhóm sẽ chia ra làm 2 file malware và benign riêng biệt vì cách thu thập dữ liệu sẽ khác nhau. Dưới đây là workflow tạo dataset benign



# THỰC NGHIỆM

## TẠO DATASET

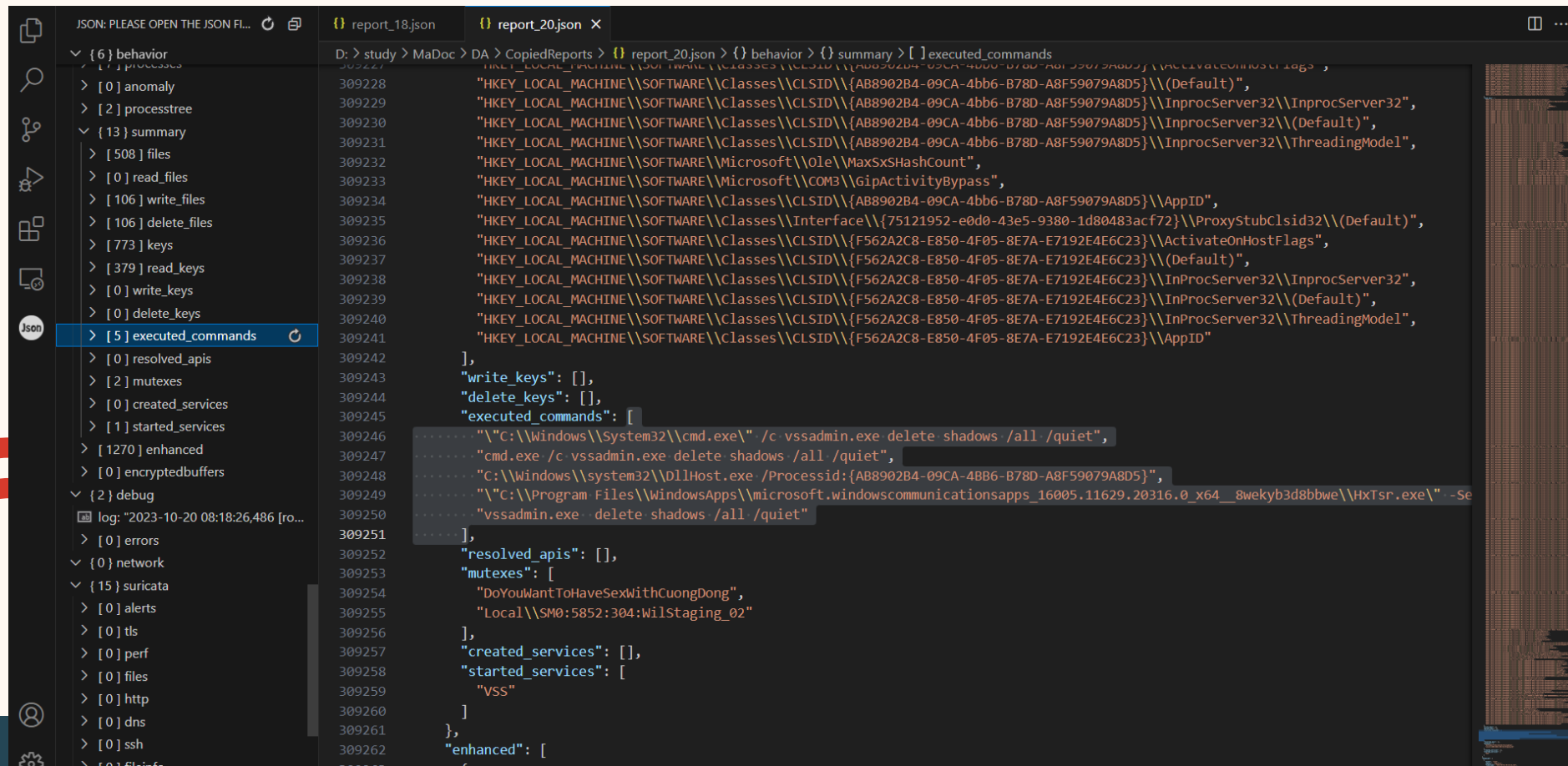
File malware thì nhóm sẽ sắp xếp theo thời gian để phục vụ cho việc phân tích time series sau này



# THỰC NGHIỆM

## TẠO DATASET

File json sau khi phân tích bằng sandbox có rất nhiều strings rải rác khắp cả file. Nhóm chỉ lấy timestamp để tạo file csv



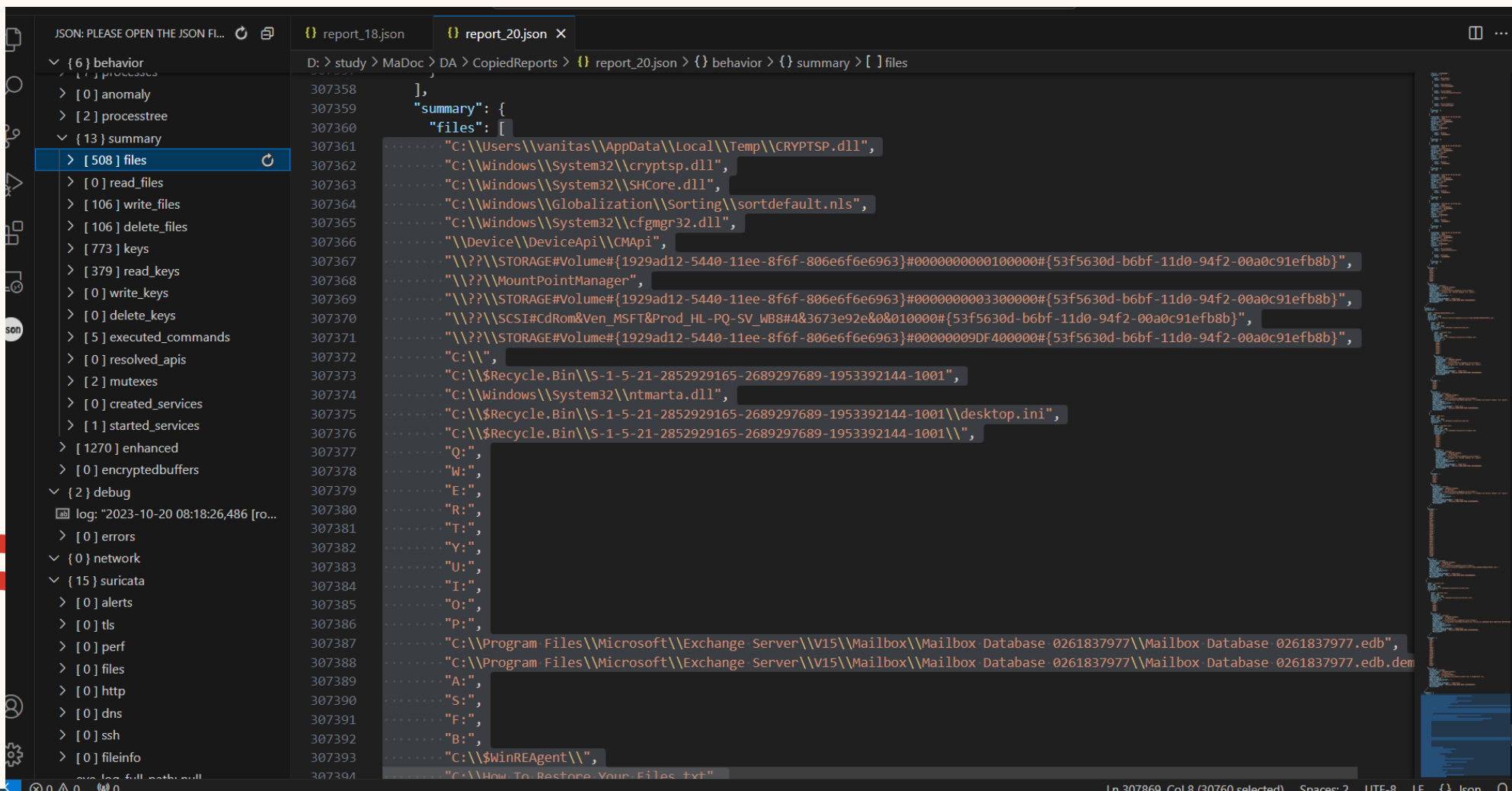
The screenshot shows a code editor with a JSON file named 'report\_20.json' open. The left sidebar shows a tree view of the JSON structure, with 'executed\_commands' selected. The main editor displays the JSON content, which is a list of objects, each containing a timestamp and a command string. The commands are related to system administration tasks, such as deleting shadows and running specific executables.

```
{
  "behavior": {
    "anomaly": [0],
    "processtree": [2],
    "summary": {
      "files": [508],
      "read_files": [0],
      "write_files": [106],
      "delete_files": [106],
      "keys": [773],
      "read_keys": [379],
      "write_keys": [0],
      "delete_keys": [0]
    },
    "executed_commands": [
      {
        "timestamp": "2023-10-20 08:18:26,486",
        "command": "cmd.exe /c vssadmin.exe delete shadows /all /quiet"
      },
      {
        "timestamp": "2023-10-20 08:18:26,486",
        "command": "cmd.exe /c vssadmin.exe delete shadows /all /quiet"
      },
      {
        "timestamp": "2023-10-20 08:18:26,486",
        "command": "C:\\Windows\\system32\\DllHost.exe /Processid:{AB8902B4-09CA-4BB6-B78D-A8F59079A8D5}"
      },
      {
        "timestamp": "2023-10-20 08:18:26,486",
        "command": "C:\\Program Files\\WindowsApps\\microsoft.windowscommunicationsapps_16005.11629.20316.0_x64_8wekyb3d8bbwe\\HxTsr.exe"
      },
      {
        "timestamp": "2023-10-20 08:18:26,486",
        "command": "vssadmin.exe delete shadows /all /quiet"
      }
    ],
    "resolved_apis": [],
    "mutexes": [
      "DoYouWantToHaveSexWithCuongDong",
      "Local\\SMO:5852:304:WilStaging_02"
    ],
    "created_services": [],
    "started_services": [
      "vss"
    ],
    "enhanced": [
      {
        "timestamp": "2023-10-20 08:18:26,486",
        "command": "cmd.exe /c vssadmin.exe delete shadows /all /quiet"
      }
    ]
  }
}
```



# TẠO DATASET

# THỰC NGHIỆM



The screenshot displays a JSON editor interface with two tabs: 'report\_18.json' and 'report\_20.json'. The 'report\_20.json' tab is active, showing a JSON structure with a 'behavior' object containing a 'summary' object, which in turn contains a 'files' array. The 'files' array lists various system files and directories, including paths like 'C:\Users\vanitas\AppData\Local\Temp\CRYPTSP.dll', 'C:\Windows\System32\cryptsp.dll', 'C:\Windows\System32\SHCore.dll', 'C:\Windows\Globalization\Sorting\sortdefault.nls', 'C:\Windows\System32\cfgmgr32.dll', '\\Device\\DeviceApi\\CMApi', '\\?\\STORAGE#Volume#{1929ad12-5440-11ee-8f6f-806e6f6e6963}#000000000100000#{53f5630d-b6bf-11d0-94f2-00a0c91efb8b}', '\\?\\MountPointManager', '\\?\\STORAGE#Volume#{1929ad12-5440-11ee-8f6f-806e6f6e6963}#0000000003300000#{53f5630d-b6bf-11d0-94f2-00a0c91efb8b}', '\\?\\SCSI#CdRom&Ven\_MSFT&Prod\_HL-PQ-SV\_WB8#4&3673e92e&0&010000#{53f5630d-b6bf-11d0-94f2-00a0c91efb8b}', '\\?\\STORAGE#Volume#{1929ad12-5440-11ee-8f6f-806e6f6e6963}#000000009DF400000#{53f5630d-b6bf-11d0-94f2-00a0c91efb8b}', 'C:\\', 'C:\\\$Recycle.Bin\\S-1-5-21-2852929165-2689297689-1953392144-1001', 'C:\\Windows\\System32\\ntmarta.dll', 'C:\\\$Recycle.Bin\\S-1-5-21-2852929165-2689297689-1953392144-1001\\desktop.ini', 'C:\\\$Recycle.Bin\\S-1-5-21-2852929165-2689297689-1953392144-1001\\', 'Q:', 'W:', 'E:', 'R:', 'T:', 'Y:', 'U:', 'I:', 'O:', 'P:', 'C:\\Program Files\\Microsoft\\Exchange Server\\V15\\Mailbox\\Mailbox Database 0261837977\\Mailbox Database 0261837977.edb', 'C:\\Program Files\\Microsoft\\Exchange Server\\V15\\Mailbox\\Mailbox Database 0261837977\\Mailbox Database 0261837977.edb', 'A:', 'S:', 'F:', 'B:', 'C:\\\$WinREAgent\\', and 'C:\\How To Restore Your Files.txt'. The left sidebar shows a tree view of the JSON structure, with the 'files' array selected. The bottom status bar indicates the current position in the file: 'In 307869, Col 8 (30760 selected) Spaces: 2 UTF-8 LF (1) json'.

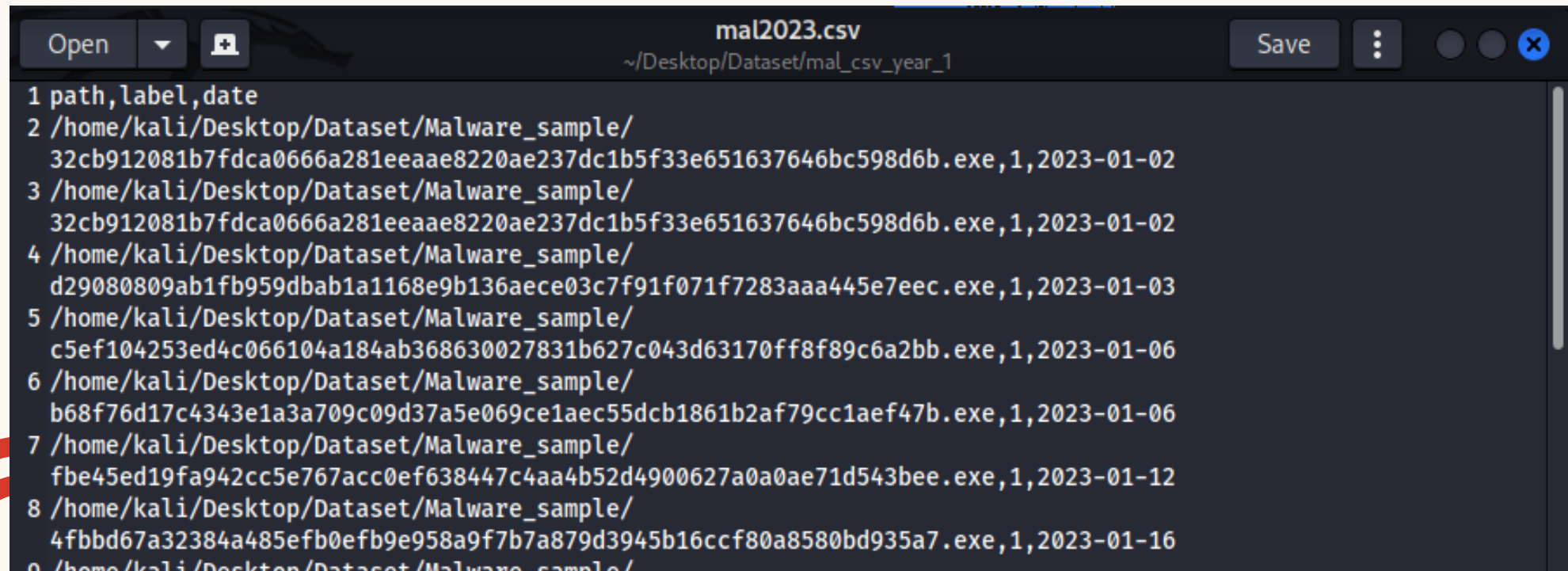
```
JSON: PLEASE OPEN THE JSON FL... report_18.json report_20.json X
D: > study > MaDoc > DA > CopiedReports > report_20.json > { } behavior > { } summary > [ ] files
307358 },
307359 "summary": {
307360 "files": [
307361 "C:\\Users\\vanitas\\AppData\\Local\\Temp\\CRYPTSP.dll",
307362 "C:\\Windows\\System32\\cryptsp.dll",
307363 "C:\\Windows\\System32\\SHCore.dll",
307364 "C:\\Windows\\Globalization\\Sorting\\sortdefault.nls",
307365 "C:\\Windows\\System32\\cfgmgr32.dll",
307366 "\\Device\\DeviceApi\\CMApi",
307367 "\\?\\STORAGE#Volume#{1929ad12-5440-11ee-8f6f-806e6f6e6963}#000000000100000#{53f5630d-b6bf-11d0-94f2-00a0c91efb8b}",
307368 "\\?\\MountPointManager",
307369 "\\?\\STORAGE#Volume#{1929ad12-5440-11ee-8f6f-806e6f6e6963}#0000000003300000#{53f5630d-b6bf-11d0-94f2-00a0c91efb8b}",
307370 "\\?\\SCSI#CdRom&Ven_MSFT&Prod_HL-PQ-SV_WB8#4&3673e92e&0&010000#{53f5630d-b6bf-11d0-94f2-00a0c91efb8b}",
307371 "\\?\\STORAGE#Volume#{1929ad12-5440-11ee-8f6f-806e6f6e6963}#000000009DF400000#{53f5630d-b6bf-11d0-94f2-00a0c91efb8b}",
307372 "C:\\",
307373 "C:\\$Recycle.Bin\\S-1-5-21-2852929165-2689297689-1953392144-1001",
307374 "C:\\Windows\\System32\\ntmarta.dll",
307375 "C:\\$Recycle.Bin\\S-1-5-21-2852929165-2689297689-1953392144-1001\\desktop.ini",
307376 "C:\\$Recycle.Bin\\S-1-5-21-2852929165-2689297689-1953392144-1001\\",
307377 "Q:",
307378 "W:",
307379 "E:",
307380 "R:",
307381 "T:",
307382 "Y:",
307383 "U:",
307384 "I:",
307385 "O:",
307386 "P:",
307387 "C:\\Program Files\\Microsoft\\Exchange Server\\V15\\Mailbox\\Mailbox Database 0261837977\\Mailbox Database 0261837977.edb",
307388 "C:\\Program Files\\Microsoft\\Exchange Server\\V15\\Mailbox\\Mailbox Database 0261837977\\Mailbox Database 0261837977.edb",
307389 "A:",
307390 "S:",
307391 "F:",
307392 "B:",
307393 "C:\\$WinREAgent\\",
307394 "C:\\How To Restore Your Files.txt"
```



# THỰC NGHIỆM

## TẠO DATASET

File csv input theo yêu cầu



The screenshot shows a text editor window titled 'mal2023.csv' with the file path '~/.Desktop/Dataset/mal\_csv\_year\_1'. The window contains a CSV file with the following content:

```
1 path,label,date
2 /home/kali/Desktop/Dataset/Malware_sample/
  32cb912081b7fdca0666a281eeaae8220ae237dc1b5f33e651637646bc598d6b.exe,1,2023-01-02
3 /home/kali/Desktop/Dataset/Malware_sample/
  32cb912081b7fdca0666a281eeaae8220ae237dc1b5f33e651637646bc598d6b.exe,1,2023-01-02
4 /home/kali/Desktop/Dataset/Malware_sample/
  d29080809ab1fb959dbab1a1168e9b136aece03c7f91f071f7283aaa445e7eec.exe,1,2023-01-03
5 /home/kali/Desktop/Dataset/Malware_sample/
  c5ef104253ed4c066104a184ab368630027831b627c043d63170ff8f89c6a2bb.exe,1,2023-01-06
6 /home/kali/Desktop/Dataset/Malware_sample/
  b68f76d17c4343e1a3a709c09d37a5e069ce1aec55dcb1861b2af79cc1aef47b.exe,1,2023-01-06
7 /home/kali/Desktop/Dataset/Malware_sample/
  fbe45ed19fa942cc5e767acc0ef638447c4aa4b52d4900627a0a0ae71d543bee.exe,1,2023-01-12
8 /home/kali/Desktop/Dataset/Malware_sample/
  4fbbd67a32384a485efb0efb9e958a9f7b7a879d3945b16ccf80a8580bd935a7.exe,1,2023-01-16
9 /home/kali/Desktop/Dataset/Malware_sample/
```

# THỰC NGHIỆM

## TẠO DATASET

File json theo đúng định dạng mà bài báo yêu cầu

```
mal_dir > {} 2ede659bfffedc8b93ce25d93e3eabe8d792512b1861316f4ab1b60cb41766e.json > [ ] strings
"Microsoft Linker", "string": "Linker: Microsoft Linker(48.0)[GUI32,admin]", "type": "Linker", "version": "48.0"}]]],
"analyze_plugin_packer": {"level": 2, "plugin_output": {"info_0": "Unusual section name found: LRT\\x04a%$S", "info_1": "Section
LRT\\x04a%$S is both writable and executable.", "info_2": "Unusual section name found:"}, "strings": ["!This program cannot be run in
DOS mode.", "a%$S", ".text", ".rsrc", "@.reloc", "UwV,V~m", "Xwvx1", "@b.$|", "3~eI", "5V!C", "l2nl", "\\w'x3", "I|In", "(By\\", ":L
[aH", "dY1b<Fh\\", "j21(T", "&yUTX", "@HLV", "Q*v\\", "NK:{D", "*q7,l", "-0.", "+pd\\", "iPe!", "[J v", "F\\t0s", "[#J.", "z(Ri", "p
{=S,E", "LB_v", ":lN\\t", "m(M|", "H^]0", ";'2\\", "%os2", "%Tr3", "nLlDj", "}%rZ", "Q:(N,", "@zP,@L", "|~Y@", "btuI", "hVLHxBLD",
"5Ms3", "UjE(", "{\\GRP", "vQ[.+zK", "g2<\\t", "t/Ckx", "}]?|7", "[NH0", "Fy3b", "6'1L", "Ki$1?", "P,6p", "CPbi", ">C_y/", "97|Z",
"QXVv", "52fp", "L@(^>71D", "BRRh", ")~Mk", "k6Ba", "<.(\\", "D?+kR", "c-mR", "RY0w", "xYwqy?", "C} 0", "q1rM", "#6n9q", "N/wi-!#1q",
"Q[jj", "}) UL", "2ns/", "GXai", "tu~*@P", "_|*w\\t", "ADJ", "/b#\\", "<bF^", "rCRM", "1jk/", "\\07m", "gl83", ">`\\v", "/ZFq",
"d:I", "btd%8", "a ~j0", ".]=\\w 7", "\\3Yd8", "p&OZH", "ahM6", "*G<=", "O&C2a", "-[lv", "oin25@1", "xFxC", "^(v(", "Co#=", "D0%SV",
"FRKUL", "z.\\tm", "M?2t#G", "m, 9C", "PSqGg", "YT^", "tRV{", "NGb[W", "p'=-", ";|X&", "ZH[&C", "j\\t5W", "2s<#", "=fGB", "6okl", "jJ_W;
", "rZ40-|", ",s8J", "5$*nL", "c;b=", "'|6.", "m:qL", "g;]H", "qS;R", "9afg", "D@;(V", "KifLm", "{@f?", "eZ\\tte", "/V/m~{", "x:V<o\\%";}
", "\\QAK", "8aaW", "Z5Lw", "0),X", "=7AU", "{j<?", "PQ};E", "D^8", "r F2", "63<z", "8mFa", "7C~B", "t51RM", "q618", "5pB<6", "v,B",
"xMu", "kHgi", "7|KZ", "2/s", "s)G1", "n)*", "e_oU*", "k$SY", "a\\t'z", "gz+=", "fLjA", "NSv", "f(+g", "P\\R\\", "x#DA", "Pz-]",
"BDz'T", "?Z15aoe", "_BYa", ">m.;P%", "{b$J", "P H6I", "(jw8", "m(rB", "YKS(*7", "15zb", "XJUUr", "/vMf\\t", "pL:k", "lZxo", "uv<-a",
"pw>Y", "UFOw", ">XLg", "K-M?", "&2Lv", "@Ttq)", "5j<~", "U$][", "eX)Z", "h4-g+Q\\t", "SIX()", "4|g", "lg?E", "OQ=La", "?~w=", ";
@%R", "gVU/", "tB&\\t", "'+ 8", "'=yU", "OY;$", "lO", "91I/", "AWB2", "qA)7", "6(i:Yi", "$q57y", "kt27", "=G]8", "VDN~", "uoy8",
"nhzf", "r17x", "T&p2", "Q#[" , "9Cc%", "@u)p", "WQ4H", "-e8:", "V~L0", "VCmZ", "f6a9", "*eSB", "5z8Tp", "RJG4", "X5vm", "fRYK",
"~1Ru", "_JIXO", "Rm(2", ">, #", "[x(fISVH", ")Y6", "E9>{", "TwEhd", "CFL{", "+b/lr", "o@l\\4", "h'q0", "C\\t8MU", "d{{F", "5pmZ",
"1at4", "B}lF", ">N(lQ*1", "[]5K1", "c&>:!", "u;&K", "Ju&OI", "gg;u", "O~89", "\\w;/", "wf %", "YT\\! ", "8Hdtat", "O\\tFZ\\", "<53t[",
"B1=K", "D[R", "b^Ry", "P?2p\\t", "vZCj", "(<'<z", "sz4)", "k*Wd", "Y2g8>", "ab>*Q", "SednI", "fkSP", "89I9", "wXIHwL", "hf[9&",
"uh#>g", "[>E", "Lro1", "O(.j", "QNY-\\t", "b|~Y|", "pLH)", "[p8&i\\t", "T7TF\\t", "JNPX", "/@8[", "6X,oi", "5k*\\t", "D?R(", "@[A*",
"hx:F-", "F_O_J", "C<wU", "x&\\g", "3d;s", "dJ(T", "6dot", "=dr9", "J)-c", "3v'F", "34(1", "eCTEGY", "jY<i", "h\\lRk", "s^LA", "ns9e",
"-N[", "t'f4", "N!r=", "%_r;", "/w:|&", "K Tp", "rAYI", "WI~5\\I", "5|:=", "bzdL", "?UP|", "Icp3", "-.CL9|", "aunH", "2FV\\v", "GDi-$,
", "P>yLpPyY", "Nu^5", "t'x", "+Zz8", "o/-", "NPT9", "k o^", "IQY5V\\", "m%o{", "IX\\Z", "RF\\8", "e+8A", "ea['> ", "lJrK", "K+;
', "VUS", "_y4%<{", "n{&%&B", "[TG^+", "5/o", "ssSH", "BRF:", "vw;sn", "Gs.AUX0E", "_8]", "3TY@e4", "CRVPQ", "\\N8|", "[Q84",
"@IV$", "~4<s", "cuqo", "&l*|", "sG~p", "R# :y", "LI#~", "wfzJ", "K-ty", "&$v[", "kbpu", "3$~t", "BI+\\", "j<lB", "A/+gJ", "60,p",
"A.|D(", "z|,8", "\\t! Y^", "Vw+m", "okb_y", "%SB[szfb", "P[1CW", "C&|Y", "ot&H", "\\l\\#", "7C$4", "tnxf", "=Q02C", "t#ri", "Y1l",
"\\1uP", "A6es", "C. $", "Q&G(pr", "QQRsM", "Redm", "F? kUN", "XwX|", "P.M)", "I g0H", "+/G", ".Z7RR", "lD *", "hWDJ", "GqPv\\",
"XkTs", "o6|i", "M&2y", "SNCT", "y_5mz:", "R36_", "0\\tD", "1 X?", "<0.", "slxV|c", "r_", ">AY", "~.S~", "F[v$", "Jl+^", "sPk",
";Q=]g&", "%8^s", "n#7|", "_0?T@", "NGR", "8-/1", "j3us", "k$539", "mEa.", "t$DQ", "@1Ao", "H 'lE", "3If>", "Hb[M", "jw)T", "2gm",
"23)K", "o9QH", "Hxyn", ")1R#", "+ Yr", "+-()", "sZ (", "<^E<", "ps\\Z", "jM", "%&8S", "3Z ?", "-Z a", "CGa8y", "%&8a", "%&8G",
"=Za8", "1CZ", "Bkz", "%&8H", "nmZ", "lFa8", "mkz", "Hfa8", "&2iDz", "[+a8b", "A_LKZ", "wa8:", "UZA8", "aT O", "Z ZBX^a8o",
"%Z =", "hZ x", "9a8)", "Z v1", "2fvZ l", "lZ &.", "a8i", "vZ ZA", "lZ \\tR", "sZ l7", "K) Z", "H8", "H7_g5g", "CorExeMain"
```

# THỰC NGHIỆM

## TẠO DATASET

File csv bao gồm các printable strings được trích xuất từ các file json

```
out_put > unknown2 > unknown2.csv
```

```
1  !This program cannot be run in DOS mode.,Rich,.text,`.rdata,@.data,.rsrc,^[,SUVW,^[,WUhx,^[Y,D$Df,\$Hf=,"\",u~,D$dQ,D$hR,D$dj,L$<R
2  ,!This program cannot be run in DOS mode.,Rich,.rdata,@.data,.idata,.rsrc,@.reloc,Microsoft Windows XP ,Microsoft Windows Millenium
3  !This program cannot be run in DOS mode.,qHy'q&,qHy%q!,qHy#q&,qHy$q&,qRich',.text,`.rdata,@.data,.rsrc,h`!P,hP!P,h 2P,h0!P,h 2P,hE!P,uPu
4  !This program cannot be run in DOS mode.,.text,`.rsrc,@.reloc,BSJB,v4.0.30319,#Strings,#GUID,#Blob,<Module>,tmpF408.tmp,Program,mscorlib
5  !This program cannot be run in DOS mode.,.text,`.rsrc,@.reloc,BSJB,v4.0.30319,#Strings,#GUID,#Blob,<Module>,tmpA6F3.tmp,Program,mscorlib
6  !This program cannot be run in DOS mode.,.text,`.rsrc,@.reloc,BSJB,v4.0.30319,#Strings,#GUID,#Blob,<Module>,tmpCE85.tmp,Program,mscorlib
7  !This program cannot be run in DOS mode.,.text,`.rsrc,@.reloc,BSJB,v4.0.30319,#Strings,#GUID,#Blob,<Module>,tmp33D2.tmp,Program,mscorlib
8  This program must be run under Win32,.text,`.itext,`.data,.bss,.idata,.didata,.edata,@.tls,.rdata,@.reloc,B.rsrc,Boolean,False,True,Syst
9  !This program cannot be run in DOS mode.,.text,`.rsrc,@.reloc,BSJB,v4.0.30319,#Strings,#GUID,#Blob,<Module>,tmpE3B6.tmp,Program,mscorlib
10 !This program cannot be run in DOS mode.,.text,`.rsrc,@.reloc,BSJB,v4.0.30319,#Strings,#GUID,#Blob,<Module>,tmpF5C1.tmp,Program,mscorlib
11 !This program cannot be run in DOS mode.,BRich+S,.text,`.CODE,`.rdata,@.data,DATA,.rsrc,XAQS,h*RA,t$:a,XAQS,XAQS,XAQS,uRFGht,XAQS,XAQS,XA
12 !This program cannot be run in DOS mode.,hRich,.text,`.rdata,@.data,.rsrc,@.reloc,Y_^[,j Y+,j Y+,Y_^[,8csm,%0;A,Y__^[,%4;A,Genu,5ineI,5n
13 !This program cannot be run in DOS mode.,Rich,.text,`.rdata,@.data,.rsrc,@.reloc,D$(j,L$(QRPV,T$ j,D$(PQRV,D$dj,L$0Q,T$DR,D$hD,D$0j,]_^3
14 !This program cannot be run in DOS mode.,Rich,.text,`.rdata,@.data,.rsrc,49t$,TVWj,PVWh,tE9u,PVWw,SVWjcf,X_^[,X_^[,^t19,QPPh,tXVP,w>WV,X
15 !This program cannot be run in DOS mode.,Rich,.text,`.rdata,@.data,.rsrc,kvgahfk,xwnvbpx,cgfunhb,^[,SUVW,^[,WUhx,^[Y,D$Df,\$Hf=,"\",u~,D$dQ,D$hR,D$dj,L$<R
16 !This program cannot be run in DOS mode.,Rich,.text,`.rdata,@.data,.rsrc,@.reloc,8SVW,Y_^[,8csm,~pjCXf,uPVWh,YYht,<v5h,Ot A,SSSSS,jdhp
17 !This program cannot be run in DOS mode.,Rich,.text,`.rdata,@.data,.rsrc,@.reloc,8SVW,Y_^[,8csm,~pjCXf,uPVWh,YYht,<v5h,Ot A,SSSSS,jdhp
```

# THỰC NGHIỆM

## ĐẶC ĐIỂM FILE BENIGN

File benign chứa các chuỗi như khai báo section ".text", ".rdata", ".data", ".pdata", ".rsrc", ".reloc" ; các API xử lý tiến trình như "GetCurrentProcessId", "GetCurrentThreadId", "OpenProcess", "GetLastError", "CreateFileA"; chuỗi liên quan đến thư viện và xử lý lỗi như "bad allocation", "bad exception", "invalid string position", "GetLastError", "TerminateProcess" ; hay các chứng chỉ "Microsoft Code Signing PCA", <http://www.microsoft.com>,... Nói chung nhìn qua ta thấy đây là 1 file thực thi bình thường chứ không có gì bất thường



# THỰC NGHIỆM

## ĐẶC ĐIỂM FILE BENIGN

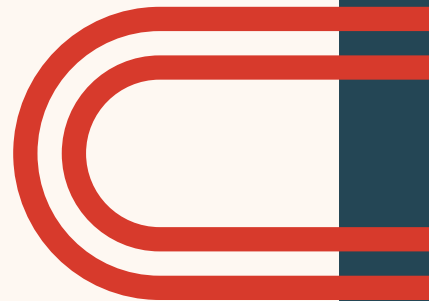
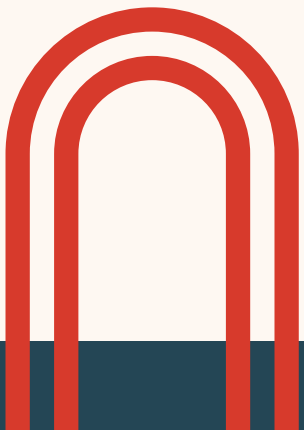
```
"D:\\a\\_work\\1\\s\\artifacts\\obj\\coreclr\\windows.x64.Release\\debug\\createdump\\createdump.pdb", "GCTL", ".text$di", ".text$mn",
".text$mn$00", ".text$mn$21", ".text$x", ".idata$5", ".00cfg", ".CRT$XCA", ".CRT$XCAA", ".CRT$XCU", ".CRT$XCZ", ".CRT$XIA", ".
CRT$XIAA", ".CRT$XIAC", ".CRT$XIZ", ".CRT$XPA", ".CRT$XPZ", ".CRT$XTA", ".CRT$XTZ", ".gehcont", ".gfids", ".rdata", ".rdata$00", ".
rdata$r", ".rdata$voltmd", ".rdata$zzzdbg", ".rtc$IAA", ".rtc$IZZ", ".rtc$TAA", ".rtc$TZZ", ".xdata", ".xdata$x", ".idata$2", ".
idata$3", ".idata$4", ".idata$6", ".data", ".data$r", ".data$rs", ".bss", ".pdata", "_RDATA", ".rsrc$01", ".rsrc$02", "GetTempPathA",
"QueryPerformanceFrequency", "QueryPerformanceCounter", "GetCurrentProcess", "OpenProcess", "GetLastError", "CreateFileA",
"CloseHandle", "K32GetModuleBaseNameA", "LocalFree", "FormatMessageA", "KERNEL32.dll", "MiniDumpWriteDump", "dbghelp.dll", "WS2_32.
dll", "NtQueryInformationProcess", "RtlCaptureContext", "RtlLookupFunctionEntry", "RtlVirtualUnwind", "ntdll.dll", "fopen",
"_acrt_iob_func", "fflush", "fclose", "__stdio_common_vfprintf", "_errno", "strcat_s", "strerror", "_time64",
"_invalid_parameter_noinfo_noreturn", "__stdio_common_vsnprintf_s", "_callnewh", "malloc", "_seh_filter_exe", "_set_app_type",
"_setusermatherr", "_configure_narrow_argv", "_initialize_narrow_environment", "_get_initial_narrow_environment", "_initterm",
"_initterm_e", "exit", "_exit", "_set_fmode", "_p__argc", "_p__argv", "_cexit", "_c_exit",
"_register_thread_local_exe_atexit_callback", "_configthreadlocale", "_set_new_mode", "_p__commode", "free",
"_initialize_onexit_table", "_register_onexit_function", "_crt_atexit", "terminate", "api-ms-win-crt-stdio-l1-1-0.dll",
"api-ms-win-crt-runtime-l1-1-0.dll", "api-ms-win-crt-string-l1-1-0.dll", "api-ms-win-crt-time-l1-1-0.dll", "api-ms-win-crt-heap-l1-1-0.
dll", "api-ms-win-crt-math-l1-1-0.dll", "api-ms-win-crt-locale-l1-1-0.dll", "UnhandledExceptionFilter", "SetUnhandledExceptionFilter",
"TerminateProcess", "IsProcessorFeaturePresent", "GetCurrentProcessId", "GetCurrentThreadId", "GetSystemTimeAsFileTime",
"InitializeSListHead", "IsDebuggerPresent", "GetModuleHandleW", "RtlUnwindEx", "RtlPcToFileHeader", "RaiseException", "SetLastError",
```

# THỰC NGHIỆM

## ĐẶC ĐIỂM FILE MALWARE

Các file malware có nhiều chuỗi trùng với file benign vì nó được chèn vào file benign nhưng vẫn có nhiều dấu hiệu nhận biết.

Đầu tiên là các file malware thường chứa các chuỗi "Encrypt", "Decrypt", "Crypto" nhằm mục đích mã hoá các thông tin của máy nạn nhân



notification of the...  
"encryptionAesRsa", "encryptedFileExtension", "checkSpread", "spreadName", "checkStartupFolder", "checkSleep", "sleepTextbox", "base64Image", "appMutexStartup", "droppedMessageTextbox", "checkAdminPrivilege", "checkdeleteShadowCopies", "checkdisableRecoveryMode", "checkdeleteBackupCatalog", "disableTaskManager", "appMutexStartup2", "appMutex2", "staticSplit", "appMutex", "System.Text.RegularExpressions", "Regex", "appMutexRegex", "System.Collections.Generic", "List`1", "messages", "validExtensions", "SystemParametersInfo", "Main", "sleepOutOfTempFolder", "AlreadyRunning", "Random", "random", "RandomString", "RandomStringForExtension", "Base64EncodeString", "encryptDirectory", "checkDirContains", "rsaKey", "CreatePassword", "AES\_Encrypt", "AES\_Encrypt\_Small", "AES\_Encrypt\_Large", "GenerateRandomSalt", "RSA\_Encrypt", "lookForDirectories", "copyRoaming", "copyResistForAdmin", "addLinkToStartup", "addAndOpenNote", "isOver", "registryStartup", "spreadIt", "runCommand", "deleteShadowCopies", "disableRecoveryMode", "deleteBackupCatalog", "DisableTaskManager", "SetWallpaper", ".ctor", "AddClipboardFormatListener", "SetParent", "IntPtr", "currentClipboard", "RegexResult", "Message", "WndProc", "CreateParams", "get\_CreateParams", "GetText", "SetText", "action", "uParam", "vParam", "winIni", "args", "length", "plainText", "location", "directory", "inputFile", "password", "keyRSA", "passwordBytes", "lengthBytes", "textToEncrypt", "publicKeyString", "commands", "base64", "System.Runtime.InteropServices", "MarshalAsAttribute", "UnmanagedType", "hwnd", "hwndChild", "hwndNewParent", "pattern", "System.Runtime.CompilerServices", "CompilationRelaxationsAttribute", "RuntimeCompatibilityAttribute", "DllImportAttribute", "user32.dll", "<Main>b\_0", "System.Threading", "ThreadStart", "CS\$<>9\_CachedAnonymousMethodDelegate2", "CompilerGeneratedAttribute", "<Main>b\_1", "CS\$<>9\_CachedAnonymousMethodDelegate3", "Thread", "Start", "Environment", "Exit", "Application", "System.Reflection", "Assembly", "GetEntryAssembly", "get\_Location", "System.IO", "Path", "GetDirectoryName", "SpecialFolder", "GetFolderPath", "String", "op\_Inequality", "Sleep", "System.Diagnostics", "Process", "GetProcesses", "GetCurrentProcess", "ProcessModuleCollection", "get\_Modules", "ProcessModule", "get\_Item", "get\_FileName", "GetExecutingAssembly", "op\_Equality", "get\_Id", "Exception", "System.Text", "StringBuilder", "get\_Length", "Next", "get\_Chars", "Append", "ToString", "Encoding", "get\_UTF8", "GetBytes", "Convert", "ToBase64String", "<>c\_DisplayClass5", "extension", "<encryptDirectory>b\_4", "ToLower", "Directory", "GetFiles", "GetExtension", "GetFileName", "Predicate`1", "Array", "Exists", "FileInfo", "FileSystemInfo", "FileAttributes", "set\_Attributes", "Concat", "File", "IEnumerable`1", "WriteAllLines", "GetDirectories", "DirectoryInfo", "get\_Attributes", "Contains", "AppendLine", "Byte", "<PrivateImplementationDetails>{6E3CDE17-2B68-4BCF-88D7-566471C4E750}", "\$method0x600000d-1", "RuntimeHelpers", "RuntimeFieldHandle", "InitializeArray", "FileStream", "FileMode", "System.Security.Cryptography", "RijndaelManaged", "SymmetricAlgorithm", "set\_KeySize", "set\_BlockSize", "PaddingMode", "set\_Padding", "Rfc2898DeriveBytes", "get\_KeySize", "DeriveBytes", "set\_Key", "get\_BlockSize", "set\_IV", "CipherMode", "set\_Mode", "Stream", "Write", "ICryptoTransform", "CreateEncryptor", "CryptoStream", "CryptoStreamMode", "CopyTo", "Flush", "Close", "FileAccess", "StreamWriter", "TextWriter", "IDisposable", "Dispose", "WriteAllText", "Delete", "get\_ASCII", "\$method0x600000e-1", "FileShare", "SetLength", "RNGCryptoServiceProvider", "RandomNumberGenerator", "RSACryptoServiceProvider", "AsymmetricAlgorithm", "FromXmlString", "Encrypt", "set\_PersistKeyInCsp", "<>c\_DisplayClass8", "dirName", "<lookForDirectories>b\_7", "DriveInfo", "GetDrives", "get\_SystemDirectory", "GetPathRoot", "get\_Name", "AppDomain",

# THỰC NGHIỆM

## ĐẶC ĐIỂM FILE MALWARE

Nó sẽ tham chiếu các registry key để chỉnh sửa hoặc thêm mới các giá trị khi cài đặt trong registry. Thông thường, các hacker sẽ sử dụng các cơ chế bền bỉ (persistence mechanisms) thông qua registry autorun, khiến mã độc tự động thực thi mỗi khi hệ thống được khởi động lại. Vậy nên cái chuỗi thường xuất hiện là RegCreateKeyEx(), RegOpenKeyEx(), RegSetValueEx(), RegDeleteKeyEx(), RegGetValue()

```
ut > 2022 > {} 0a425ea8985ae8d03a80943d8060891f38a3576a3575fb0e937db2797e3af198.json > [ ] strings > 345
"SVW3", "<At/", "ZYYd", "ZYYd", "Uh\"o@", "ZYYd", "h)o@", "SVW3", "ZYYd", "_^[YY
q@", "ZYYd", "h-q@", "ZYYd", "SVW3", "Zu|3", "ZYYd", "\\PROGRA~1\\", "QQQQQQSVW",
"hcv@", "_^[YY]", "ZYYd", "ZYYd", "ZYYd", "open", "QQQQQQS3", "ZYYd", "QQQQQQ", "ZYYd", "ZYYd", "UhJ}@", "ZYYd", "hQ}@", "ZYYd",
"QQQQQQSV", "Uhu~@", "ZYYd", "h|~@", "ZYYd", "ZYYd", "ZYYd", "ZYYd", "Error", "Runtime error at 00000000", "0123456789ABCDEF",
"kernel32.dll", "DeleteCriticalSection", "LeaveCriticalSection", "EnterCriticalSection", "InitializeCriticalSection", "VirtualFree",
"VirtualAlloc", "LocalFree", "LocalAlloc", "GetVersion", "GetCurrentThreadId", "GetThreadLocale", "GetStartupInfoA", "GetLocaleInfoA",
"GetCommandLineA", "FreeLibrary", "ExitProcess", "WriteFile", "UnhandledExceptionFilter", "RtlUnwind", "RaiseException",
"GetStdHandle", "user32.dll", "GetKeyboardType", "MessageBoxA", "advapi32.dll", "RegQueryValueExA", "RegOpenKeyExA", "RegCloseKey",
"oleaut32.dll", "SysFreeString", "SysReAllocStringLen", "kernel32.dll", "TlsSetValue", "TlsGetValue", "LocalAlloc",
"GetModuleHandleA", "advapi32.dll", "RegSetValueExA", "RegOpenKeyExA", "RegCloseKey", "kernel32.dll", "WriteFile", "WinExec",
"SetFilePointer", "SetFileAttributesA", "SetEndOfFile", "SetCurrentDirectoryA", "ReleaseMutex", "ReadFile", "GetWindowsDirectoryA",
```

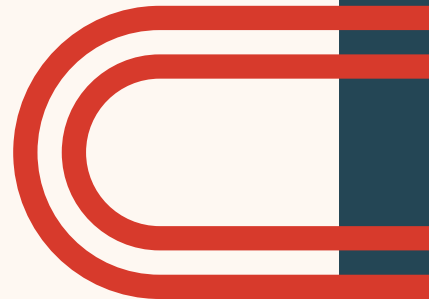
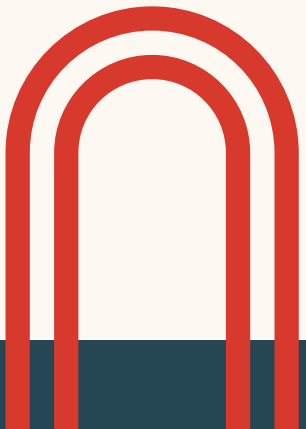


# TRIỂN KHAI PHƯƠNG PHÁP

## ĐẶC ĐIỂM FILE MALWARE

Ở đây ta thấy file này được viết bằng ngôn ngữ "Delphi" chứ không viết bằng C++ như các file benign thông thường. Delphi được sử dụng phổ biến để viết mã độc vì nó có thể tạo các chương trình mới dễ dàng, nhanh chóng và tận dụng được các API của Windows.

under win32 , CODE , DATA , .ldata , .t1s , .pdata , P.reloc , P.rsrc , %QA , %(QA , %\$QA , %QA , %8QA , %HQA , %DQA  
"%@QA", "%TQA", "%PQA", "u:hd", "SVWUQ", "Z]\_^[" , "SVWU", "YZ]\_^[" , "SVWU", " ]\_^[" , "SVWU", "w;;t\$", " ]\_^[" , "SVWU", " ]\_^[" , "SVWUQ",  
"Z]\_^[" , "SVWU", "YZ]\_^[" , "SVWU", "uW;{" , "u:;{" , " ]\_^[" , "ZYYd", "ZYYd", "SVWU", " ]\_^[" , "YZ^[" , "SVWU", " ]\_^[" , "ZYYd", " \_^[YY]",  
"QSVW", "ZYYd", " \_^[Y]", "SVWU", "\$;L\$", "\$)D\$", "YZ]\_^[" , "QSVW", "Uh5\$@", "ZYYd", "h<\$@", " \_^[Y]", "r/f=", "w)f%", "uEnt", "u0Nt",  
"u%Nt", "\tw%9", "~KxI()", "2 \_^[", "%4QA", "ZYYd", "SOFTWARE\\Borland\\Delphi\\RTL", "FPUMaskValue", "PPRTj", "YYZX", "YZXtp", "vWUD"  
"SPRQ", "T\$(j", "SVWU", " ]\_^[" , " ]\_^[" , "d\$,1", ",t\\=", "t=HtN", "r6t0", "t.Ht", "ZYYd", " \_^[", "Uh2.@", "ZYYd", " \_^[", "SVWU", " ]  
[" , "; \_^[", "t!R:", "SVWRP", "Z \_^[X", "uXJt", "uAJt", "u:Jt", "It2S", "t&J|", "N|\*9", "t1SVW", "; \_^[", "PSVW", " \_^[X", " \_^[X", "SVWU"

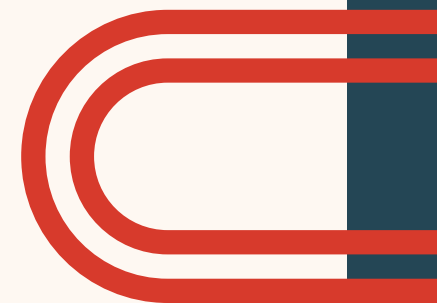
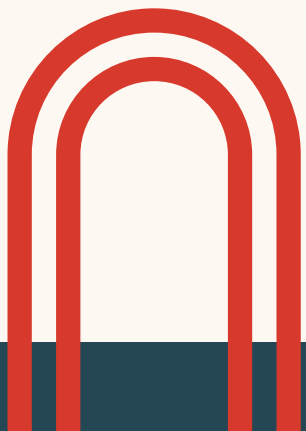


# THỰC NGHIỆM

## ĐẶC ĐIỂM FILE MALWARE

Mã độc sử dụng phương pháp anti debugger hoặc anti vm sẽ có các chuỗi như "IsDebuggerPresent", "DebugBreak", "CreateFileA/W", "CreateFileMappingA/W"

```
"OutputDebugStringA", "GetUserDefaultLCID", "InitializeCriticalSectionEx", "DecodePointer", "DeleteCriticalSection", "VirtualAlloc",  
"IsDebuggerPresent", "DebugBreak", "InitializeCriticalSection", "EnterCriticalSection", "LeaveCriticalSection", "GetModuleFileNameW",  
"GetConsoleScreenBufferInfo", "SetConsoleTextAttribute", "LoadLibraryA", "SearchPathW", "GetCurrentProcessId", "GetWindowsDirectoryW",  
"VirtualFree", "Sleep", "GetModuleHandleA", "GetCurrentProcess", "SetEnvironmentVariableW", "ExpandEnvironmentStringsW",  
"FreeLibrary", "GlobalMemoryStatusEx", "GetLocalTime", "RaiseException", "CreateFileA", "GetFileSize", "MapViewOfFile",  
"UnmapViewOfFile", "GetModuleFileNameA", "CreateFileMappingA", "LoadLibraryExA", "FindFirstFileW", "MapViewOfFileEx", "LocalFree",  
"FormatMessageA", "EnumResourceLanguagesA", "BeginUpdateResourceA", "UpdateResourceA", "EndUpdateResourceA", "SetLastError",  
"SetFileAttributesW", "GetCurrentThreadId", "CreateFileMappingW", "LoadLibraryExW", "DeleteFileW", "GetFileType", "SetEndOfFile",  
"SetFilePointerEx", "DeviceIoControl", "InitializeCriticalSectionAndSpinCount", "InitializeSRWLock", "ReleaseSRWLockExclusive",
```



# THỰC NGHIỆM

## ĐẶC ĐIỂM FILE MALWARE

Mã độc sử dụng kỹ thuật process hollowing để unmap các đoạn code hợp lệ khỏi bộ nhớ của tiến trình và tải lên các đoạn code độc hại

```
DNS response unrecognized address get adapters addresses unexpected network: 37252902984619140625 invalid request code bad font file format is
a named type file key has been revoked connection timed
ut CreateProcessAsUserW CryptAcquireContextW CertOpenSystemStoreW GetCurrentDirectoryW GetFileAttributesExW SetCurrentDirectoryW SetHandleInt
KGetProcessTimes DuplicateHandle negative offset is not defined 476837158203125 advertise error key has expired network is down
found no such
s GetAdaptersInfo CreateHardLinkW DeviceIoControl FlushViewOfFile GetCommandLineW GetStartupInfoW Process32FirstW UnmapViewOfFile
#v\lan_tciCfMgr32.dll setupapi.dll wintrust.dll wtsapi32.
dllOpenServiceW ReportEventW CreateMutexW GetProcessId LoadResource LockResource ReleaseMutex ResumeThread SetErrorMode SetStdHandle Thread32New
VirtualAlloc NtCreateFile CoCreateGuid AMDIsbetter! AuthenticAMD CentaurHauls GenuineIntel Transmeta CPU GenuineTMMx86 Geode by NSC VIA VIA VIA
KVM KVM KVM KVM Microsoft Hv VMWare VMWare Xen VMM Xen VMM bhyve bhyve Hygon GenuineVortex86 SoC SiS SiS SiS Rise Rise Rise Genuine RDCnot
closed CM_MapCr10Win32Err CloseServiceHandle CreateWellKnownSid GetSidSubAuthority MakeSelfRelativeDllQueryServiceStatus CertGetNameStringW Crypt
UnprotectData PFXImport CertStoreGetBestInterfaceEx ClosePseudoConsole GetCurrentThreadId GetModuleHandleExW GetVolumePathNameW RemovedDllDir
ectory TerminateJobObject WriteProcessMemory EnumProcessModules GetModuleBaseNameW Ed25519-Dilithium2 wrong input length n must be
positive tag:yaml.org,2002:snowflake-client %s proxy test failure: bad flag syntax: %s reflect.Value.IsNil reflect.Value.Float criterion too
ISA110C, SWATCH011111111111, SUSPEND111111111111, SETWAITABLETIME1, SETCOMMANDED EXCEPTION FILTER, SETPROCESSPRIORITY BOOST,
"SetEvent", "SetErrorMode", "SetConsoleCtrlHandler", "ResumeThread", "RaiseFailFastException", "PostQueuedCompletionStatus",
"LoadLibraryW", "LoadLibraryExW", "SetThreadContext", "GetThreadContext", "GetSystemInfo", "GetSystemDirectoryA", "GetStdHandle",
"GetQueuedCompletionStatusEx", "GetProcessAffinityMask", "GetProcAddress", "GetErrorMode", "GetEnvironmentStringsW".
```

# THỰC NGHIỆM

Classifier	Parameter	Optimum value	
		<b>Doc2Vec</b>	<b>LSI</b>
	kernel	rbf	rbf
<b>SVM</b>	C	100	100
	gamma	0.01	0.1
<b>RF</b>	n_estimators	392	442
	n_jobs	14	32
	max_depth	11	3
<b>XGB</b>	min_child_weight	11	5
	subsample	0.5	0.8
	colsample_tree	0.7	0.6
<b>MLP</b>	optimizer	Adam	Adam
	epoch	40	40
<b>CNN</b>	optimizer	Adam	Adam
	epoch	40	40

# THỰC NGHIỆM

## TẠO DATASET

Kết quả nhóm tạo được các tập dataset với số lượng sample như sau:

- ❑ Malware train + test = 290 + 211
- ❑ Benign train: 255
- ❑ Malware unknown1 + unknown2: 409 + 230
- ❑ Benign unknown1 + unknown2: 85 + 85

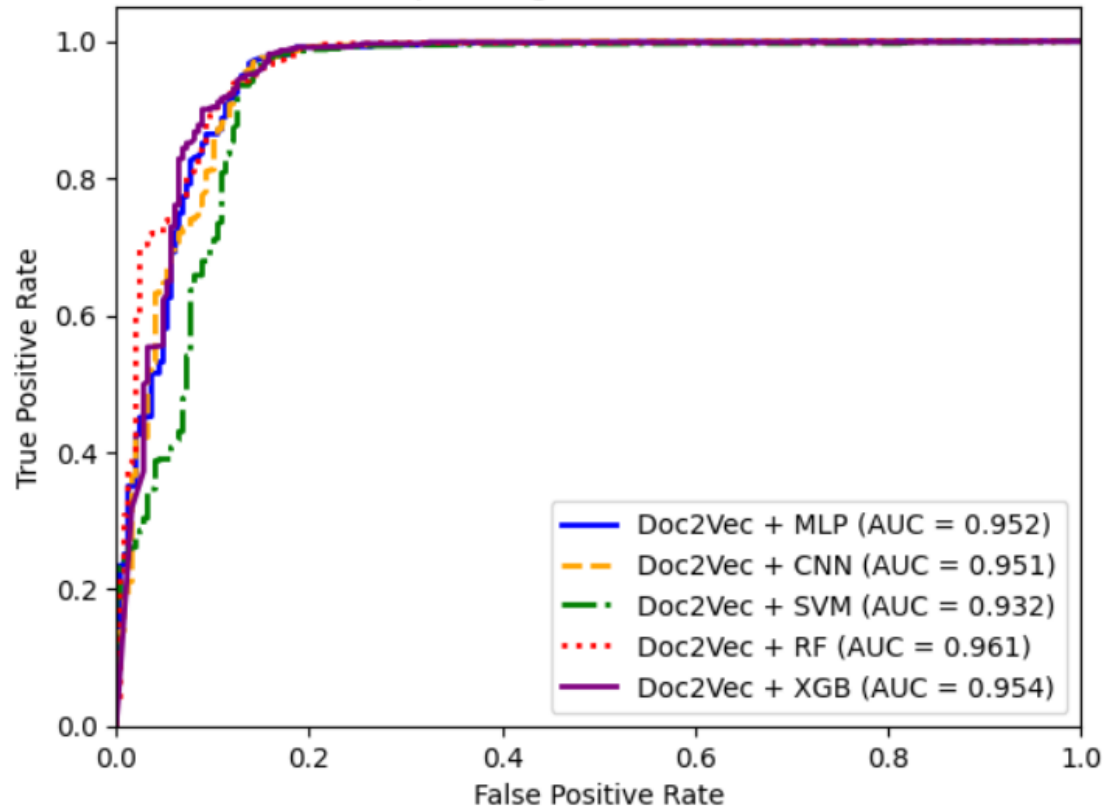
*Trong đó, malware unknown1 có **timestamp < unknown2** (sử dụng cả 2 tập để kiểm tra khả năng phát hiện malware mới cho các mô hình)*

# THỰC NGHIỆM

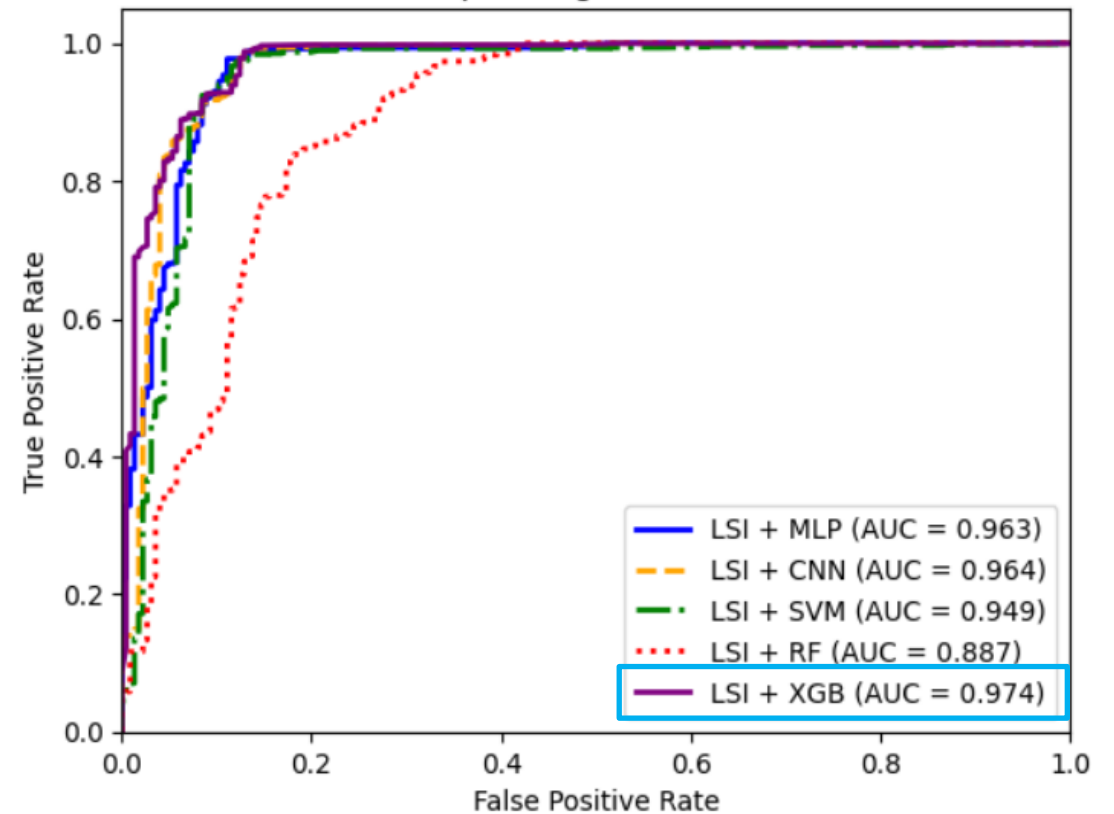
	SVM		RF		XGB		MLP		CNN	
	Doc2Vec	LSI	Doc2Vec	LSI	Doc2Vec	LSI	Doc2Vec	LSI	Doc2Vec	LSI
Accuracy	0.942	0.956	0.947	0.907	0.951	0.966	0.948	0.960	0.949	0.964
Precision	0.942	0.955	0.949	0.917	0.950	0.967	0.948	0.960	0.949	0.963
Recall	0.942	0.956	0.947	0.907	0.951	0.966	0.948	0.960	0.949	0.964
F1-score	0.940	0.956	0.944	0.895	0.950	0.966	0.948	0.960	0.948	0.963

# THỰC NGHIỆM

Receiver Operating Characteristic - Doc2Vec



Receiver Operating Characteristic - LSI



# THỰC NGHIỆM

	SVM		RF		XGB		MLP		CNN	
	Doc2Vec	LSI	Doc2Vec	LSI	Doc2Vec	LSI	Doc2Vec	LSI	Doc2Vec	LSI
unk1	0.620	0.754	0.688	0.732	0.678	0.778	0.584	0.756	0.686	0.785
unk2	0.628	0.606	0.679	0.672	0.650	0.681	0.610	0.710	0.652	0.685

DETECTION RATE (RECALL)

KIỂM TRA TRÊN MALWARE MỚI





# NỘI DUNG BÁO CÁO

1  
GIỚI THIỆU  
ĐỀ TÀI

3  
TRIỂN KHAI  
PHƯƠNG PHÁP

2  
PHƯƠNG PHÁP  
NGHIÊN CỨU

4  
SO SÁNH  
VÀ ĐÁNH GIÁ

# SO SÁNH VÀ ĐÁNH GIÁ

- Có thể thấy rằng với language model LSI đưa ra được các kết quả cao hơn là Doc2Vec. Sự kết hợp tốt nhất là LSI + XGB (trong paper là LSI + SVM)
- Các kết quả trong quá trình học khá ấn tượng dù với dataset tự thu thập và tạo theo script của tác giả.
- Về phần kiểm tra khả năng phát hiện trên tệp malware mới, chưa từng được training thì kết quả thu được không được cao như paper (trong paper đề cập đến detection rate new malware của LSI + SVM trung bình là 0.973).
- Tệp unknown2 có timestamp mới hơn so với unknown1 cũng tương ứng với kết quả phát hiện thấp hơn hoàn toàn phù hợp với việc malware thay đổi và tiến hóa theo thời gian cũng như khả năng thích ứng của mô hình đã học còn hạn chế; cần được mở rộng thêm dataset huấn luyện theo thời gian để đưa ra những kết quả phát hiện tốt nhất

**THANK YOU**

