gera_2021_a_semi_automated_approach_for_identification_of_trends_in_android_ransomware_literature

Year

2021

Author(s)

Tanya Gera and Jaiteg Singh and Deepak Thakur and Parvez Faruki

Title

A Semi-automated Approach for Identification of Trends in Android Ransomware Literature

Venue

Machine Learning for Networking

Topic labeling

Manual

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Manual labeling

Topic labeling parameters

\

Label generation

Loading weights of the terms for every topic solution are sorted in descending order so as to give suitable labels for high loading values with help of subject field experts.

Systematic literature Analysis over Android stealth malware corpus results in identifying three core areas as Table 9.

Table 9. Core research areas

Topic no.	Topic label	Top loading terms
TS3.1	App Structure Monitoring	signature, bytecode, graph, context, dalvik, flow, permission, component, control, library, program, service, method, object, entry, event, field, code, data, path
TS3.2	App Behaviour Monitoring	kernel, privilege, escalation, policy, control, enforcement, security, exploit, memory, vulnerability, library, native, context, component, linux, mechanism, access, resource, sandbox, virtual
TS3.3	Hybrid Level Monitoring	dynamic, analysis, static, cloud, taint, application, instruction, execution, component, sensitive, bytecode, android, library, program, native, object, string, dalvik, class, event

Table 10. Ten topic solutions

Topic no	Topic label	2009–2013	2014–2019	2009–2019
TS10.1	Emulator Based Analysis	23	26	49
TS10.2	Dynamic Code Loading	19	32	51
TS10.3	High Battery Consumption	22	21	43
TS10.4	Context Monitoring	21	30	51
TS10.5	API Call Monitoring	18	31	49
TS10.6	Dalvik Byte Code Analysis	21	17	38
TS10.7	Permission Based Analysis	22	31	53
TS10.8	Classification Based on App Behavior	19	38	57
TS10.9	Graph Based Analysis	13	29	42
TS10.10	Feature Based Analysis	23	31	54

Table 11. Twenty topic solutions

Topic no	Topic label	2009–2013	2014–2019	2009–2019
TS20.1	Obfuscated Code Analysis	5	6	11
TS20.2	Privacy Leakage Monitoring	19	7	26
TS20.3	Hybrid Analysis	5	24	29
TS20.4	Pattern Assessment	9	12	21
TS20.5	Permission Based Analysis	22	20	42
TS20.6	Kernel Level Check	12	18	30
TS20.7	Signature Based Analysis	16	19	35
TS20.8	Classification Based on App Behavior	11	21	32
TS20.9	Dynamic Code Loading	8	21	29
TS20.10	Emulator Based Analysis	7	14	21
TS20.11	Taint Analysis	4	12	16
TS20.12	Graph Based Analysis	12	13	25
TS20.13	Flow Monitoring	3	9	12
TS20.14	API Call Monitoring	6	12	18
TS20.15	User Interactions	3	13	16
TS20.16	Context Monitoring	12	22	34
TS20.17	Feature Based Analysis	4	7	11
TS20.18	Dalvik Byte Code Analysis	8	15	23
TS20.19	High Battery Consumption	4	19	23
TS20.20	Text Based Analysis	4	29	33

Motivation

\

Topic modeling

LSA

Topic modeling parameters

Nr of topics: {5, 10, 20, 30}

Nr. of topics

Label

Manually assigned single or multi word labels

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper: Malware research

Dataset: Malware research

Problem statement

This study uses Latent Semantic Analysis (LSA), an information modelling technique to deduce core research areas, research trends and widely investigated areas within corpus. This work takes a large corpus of 487 research articles (published during 2009–2019) as input and produce three core research areas and thirty emerging research trends in field of stealth malwares as primary goal. LSA, a semi-automated approach is helpful in achieving a significant innovation over traditional methods of literature review and had shown great performance in many other research fields like medical, supply chain management, open street map etc. The secondary aim of this study is to investigate popular latent topics by mapping core research trends with core research areas. This study also provides prospective future directions for heading researchers.

Corpus

Origin:

Nr. of documents: 487

Details:

- published during 2009–2019

Document

Article focussing on Android stealth malware

Pre-processing

- tokenization
- removing stop words
- normalization
- stemming & lemmatizing
- character filtering.

Table 3. Sample outcomes after Pre-processing

S. no	Pre-processing steps	After pre-processing
a.)	**Tokenization** • Converts Large Chunks of text to sentences • Sentences to Words	[‘Malware’, ‘application’, ‘reads’, ‘the’, ‘unique’, ‘device’, ‘identifier’, ‘to’, ‘track’, ‘the’, ‘user’, ‘s’, ‘device’]
b.)	**Removing Stop Words** • Remove stop words • Remove Common words	[‘Malware’, ‘application’, ‘reads’, ‘unique’, ‘device’, ‘identifier’, ‘track’, ‘user’, ‘device’]
c.)	**Normalization** • Standard Formatting • Upper case to lower case • Numbers to word equivalents	Malware application reads the unique device identifier to track the user s device
d.)	**Stemming and Lemmatizing** • Reduces total number of unique words • Converts the words to their word stem • Past and Future tenses to present • Third form to first form	[‘malwar’, ‘applic’, ‘read’, ‘uniqu’, ‘devic’, ‘identifi’, ‘track’, ‘user’, ‘devic’]
e.)	**N-Character Filtering** • Words less than the length 4 were omitted	[‘malwar’, ‘applic’, ‘read’, ‘uniqu’, ‘devic’, ‘identifi’, ‘track’, ‘user’, ‘devic’]

```
@incollection{gera_2021_a_semi_automated_approach_for_identification_of_trends_in_android_ransomware_literature,
    author = {Tanya Gera and Jaiteg Singh and Deepak Thakur and Parvez Faruki},
    booktitle = {Machine Learning for Networking},
    date-added = {2023-04-28 10:42:15 +0200},
    date-modified = {2023-04-28 10:42:15 +0200},
    doi = {10.1007/978-3-030-70866-5_18},
    pages = {265--283},
    publisher = {Springer International Publishing},
    title = {A Semi-automated Approach for Identification of Trends in Android Ransomware Literature},
    url = {https://doi.org/10.1007%2F978-3-030-70866-5_18},
    year = 2021}
```


#Thesis/Papers/FS