# natural language processing in malware detection

## Any time:

Similar to the NLP process ([Teufl et. al., 2010](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1007_978-3-642-14706-7_20)) call the process Machine Language Processing (MLP). ([Teufl et. al., 2010](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1007_978-3-642-14706-7_20)) use the e-Participation analysis architecture, extract the various NLP techniques and adopt them for the malware analysis process. ([Kolosnjaji et. al., 2017](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_ijcnn.2017.7966340)) transfer the performance improvements achieved in the area of neural networks to model the execution sequences of disassembled malicious binaries. ([Jang et. al., 2017](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1145_3134600.3134635)) present a simple gradient-descent based algorithm for finding adversarial samples, which performs well in comparison to existing algorithms. ([Huang et. al., 2017](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1145_3052973.3053017)) design Gossip, a novel approach to automatically detect malicious domains based on the analysis of discussions in technical mailing lists (particularly on security-related topics) by using natural language processing and machine learning techniques. In order to reduce the manpower of feature engineering prior to the condition of not to extract pre-selected features ([Huang et. al., 2017](https://www.paperdigest.org/paper/?paper_id=arxiv-1705.04448)) develop a coloR-inspired convolutional neuRal networks (CNN)-based AndroiD malware Detection (R2-D2) system. ([Wang et. al., 2018](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_tifs.2017.2771228)) propose an effective and automatic malware detection method using the text semantics of network traffic. The approach assessed is a novel dynamic malware analysis method, which may generalize better than static analysis to newer variants ([Kim, 2018](https://www.paperdigest.org/paper/?paper_id=arxiv-1802.05412)). ([Sewak et. al., 2018](https://www.paperdigest.org/paper/?paper_id=arxiv-1809.05889)) investigate and compared one of the Deep Learning Architecture called Deep Neural Network (DNN) with the classical Random Forest (RF) machine learning algorithm for the malware classification. Similar to natural language processing ([Lu, 2019](https://www.paperdigest.org/paper/?paper_id=arxiv-1906.04593)) propose a novel and efficient approach to perform static malware analysis, which can automatically learn the opcode sequence patterns of malware. Other influential work includes ([Kolosnjaji et. al., 2016](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1007_978-3-319-50127-7_11)).

## Past 5 years:

It is vital to call for further attention regarding security threats and corresponding defensive techniques of machine learning, which motivates a comprehensive survey ([**Liu et. al., 2018**](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_access.2018.2805680)). ([**Wang et. al., 2018**](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_tifs.2017.2771228)) propose an effective and automatic malware detection method using the text semantics of network traffic. The approach assessed is a novel dynamic malware analysis method, which may generalize better than static analysis to newer variants ([**Kim, 2018**](https://www.paperdigest.org/paper/?paper_id=arxiv-1802.05412)). ([**Sewak et. al., 2018**](https://www.paperdigest.org/paper/?paper_id=arxiv-1809.05889)) investigate and compared one of the Deep Learning Architecture called Deep Neural Network (DNN) with the classical Random Forest (RF) machine learning algorithm for the malware classification. To this end ([**Redmond et. al., 2018**](https://www.paperdigest.org/paper/?paper_id=arxiv-1812.09652)) propose a joint learning approach to generating instruction embeddings that capture not only the semantics of instructions within an architecture, but also their semantic relationships across architectures. ([**Karbab et. al., 2019**](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1016_j.diin.2019.01.017)) propose, MalDy (mal die), a portable (plug and play) malware detection and family threat attribution framework using supervised machine learning techniques. To leverage the application of deep learning architectures towards cyber security ([**Vinayakumar et. al., 2019**](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1007_978-3-030-16837-7_7)) consider intrusion detection, traffic analysis and Android malware detection. ([**Zhang et. al., 2019**](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1016_j.future.2018.12.028)) propose a cloud-based malware detection system called SaaS by leveraging and marrying multiple approaches from diverse domains such as natural language processing (n-gram), image processing (GLCM), cryptography (fuzzy hash), machine learning (random forest) and complex networks. Similar to natural language processing ([**Lu, 2019**](https://www.paperdigest.org/paper/?paper_id=arxiv-1906.04593)) propose a novel and efficient approach to perform static malware analysis, which can automatically learn the opcode sequence patterns of malware. An anomaly detection method is implemented using Natural Language Processing (NLP) based on Bags of System Calls (BoSC) for learning the behavior of applications on Windows virtual machines running on Xen hypervisor ([**Peddoju et. al., 2020**](https://www.paperdigest.org/paper/?paper_id=doi.org_10.14569_ijacsa.2020.0110559)).

## Past Year:

([**Xu et. al., 2021**](https://www.paperdigest.org/paper/?paper_id=arxiv-2112.10035)) present hybrid-Flacon, a hybrid pattern Android malware detection and categorization framework. ([**Xu, 2021**](https://www.paperdigest.org/paper/?paper_id=arxiv-2112.10038)) explore an ensemble mechanism, which presents how the combination of byte-code and native-code analysis of Android applications can be efficiently used to cope with the advanced sophistication of Android malware. To this end ([**Beliaev et. al., 2022**](https://www.paperdigest.org/paper/?paper_id=arxiv-2201.09369)) propose a novel defense method that consists of "truncation" and "adversarial training". ([**Sahabandu et. al., 2022**](https://www.paperdigest.org/paper/?paper_id=arxiv-2204.06624)) propose a binary code feature extraction model to improve the accuracy and scalability of ML-based ISA identification methods. Pre-trained neural Language Models (PTLM), such as CodeBERT, are recently used in software engineering as models pre-trained on large source code corpora ([**Goel et. al., 2022**](https://www.paperdigest.org/paper/?paper_id=arxiv-2204.08653)). Only a few recent works have considered domain-independent deception. ([**Verma et. al., 2022**](https://www.paperdigest.org/paper/?paper_id=arxiv-2207.01738)) investigate domain-independent deception along four dimensions. ([**Uhlig et. al., 2022**](https://www.paperdigest.org/paper/?paper_id=arxiv-2208.11367)) propose deep learning approximate matching (DLAM), which achieves much higher accuracy in detecting anomalies in fuzzy hashes than conventional approaches. ([**Chen et. al., 2022**](https://www.paperdigest.org/paper/?paper_id=arxiv-2208.14221)) present AVMiner, an expansible malware tagging system that can mine the most vital tokens from AV labels. ([**Maniriho et. al., 2022**](https://www.paperdigest.org/paper/?paper_id=arxiv-2209.03547)) present MalBehavD-V1, a new behavioural dataset of Windows Application Programming Interface (API) calls extracted from benign and malware executable files using the dynamic analysis approach. ([**Jawad et. al., 2022**](https://www.paperdigest.org/paper/?paper_id=arxiv-2210.08331)) describe the Fake News Challenge stage #1 (FNC-1) dataset and given an overview of the competitive attempts to build a fake news detection system using the FNC-1 dataset.

# natural language processing in malware behavioral detection

## Any time

([**Kolosnjaji et. al., 2017**](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_ijcnn.2017.7966340)) transfer the performance improvements achieved in the area of neural networks to model the execution sequences of disassembled malicious binaries. ([**Karbab et. al., 2017**](https://www.paperdigest.org/paper/?paper_id=arxiv-1702.05699)) address the following question: could ([**Karbab et. al., 2017**](https://www.paperdigest.org/paper/?paper_id=arxiv-1702.05699)) generate effective fingerprints for Android malware through dynamic analysis? In order to reduce the manpower of feature engineering prior to the condition of not to extract pre-selected features ([**Huang et. al., 2017**](https://www.paperdigest.org/paper/?paper_id=arxiv-1705.04448)) develop a coloR-inspired convolutional neuRal networks (CNN)-based AndroiD malware Detection (R2-D2) system. ([**Karbab et. al., 2018**](https://www.paperdigest.org/paper/?paper_id=arxiv-1812.10327)) propose, MalDy (mal~die), a portable (plug and play) malware detection and family threat attribution framework using supervised machine learning techniques. ([**Karbab et. al., 2019**](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1016_j.diin.2019.01.017)) propose, MalDy (mal die), a portable (plug and play) malware detection and family threat attribution framework using supervised machine learning techniques. ([**Safa et. al., 2019**](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_iwcmc.2019.8766515)) benchmark deep learning architectures composed of recurrent and convolutional neural networks. ([**Poudyal et. al., 2021**](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_access.2021.3109260)) propose a deep inspection approach for multi-level profiling of crypto-ransomware, which captures the distinct features at Dynamic link library, function call, and assembly levels. ([**Parildi et. al., 2021**](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1007_s00521-021-05861-7)) present an alternative method for malware detection, which makes use of assembly opcode sequences obtained during runtime. Other influential work includes ([**Kolosnjaji et. al., 2016**](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1007_978-3-319-50127-7_11)).

## Past 5 years

([Wang et. al., 2018](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_tifs.2017.2771228)) propose an effective and automatic malware detection method using the text semantics of network traffic. The approach assessed is a novel dynamic malware analysis method, which may generalize better than static analysis to newer variants ([Kim, 2018](https://www.paperdigest.org/paper/?paper_id=arxiv-1802.05412)). ([Sewak et. al., 2018](https://www.paperdigest.org/paper/?paper_id=arxiv-1809.05889)) investigate and compared one of the Deep Learning Architecture called Deep Neural Network (DNN) with the classical Random Forest (RF) machine learning algorithm for the malware classification. ([Karbab et. al., 2018](https://www.paperdigest.org/paper/?paper_id=arxiv-1812.10327)) propose, MalDy (mal~die), a portable (plug and play) malware detection and family threat attribution framework using supervised machine learning techniques. ([Karbab et. al., 2019](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1016_j.diin.2019.01.017)) propose, MalDy (mal die), a portable (plug and play) malware detection and family threat attribution framework using supervised machine learning techniques. ([Safa et. al., 2019](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_iwcmc.2019.8766515)) benchmark deep learning architectures composed of recurrent and convolutional neural networks. Similar to natural language processing ([Lu, 2019](https://www.paperdigest.org/paper/?paper_id=arxiv-1906.04593)) propose a novel and efficient approach to perform static malware analysis, which can automatically learn the opcode sequence patterns of malware. MDTA is the best suitable and manageable approach for analyzing behavioral reports using a machine learning algorithm for providing security measures to identify malware without the intervention of the investigator ([Vanjire et. al., 2021](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1007_978-981-16-1866-6_10)). ([Poudyal et. al., 2021](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_access.2021.3109260)) propose a deep inspection approach for multi-level profiling of crypto-ransomware, which captures the distinct features at Dynamic link library, function call, and assembly levels. ([Parildi et. al., 2021](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1007_s00521-021-05861-7)) present an alternative method for malware detection, which makes use of assembly opcode sequences obtained during runtime.

## Past year

([Xu et. al., 2021](https://www.paperdigest.org/paper/?paper_id=arxiv-2112.10035)) present hybrid-Flacon, a hybrid pattern Android malware detection and categorization framework. ([Xu, 2021](https://www.paperdigest.org/paper/?paper_id=arxiv-2112.10038)) explore an ensemble mechanism, which presents how the combination of byte-code and native-code analysis of Android applications can be efficiently used to cope with the advanced sophistication of Android malware. To this end ([Beliaev et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2201.09369)) propose a novel defense method that consists of "truncation" and "adversarial training". ([Sahabandu et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2204.06624)) propose a binary code feature extraction model to improve the accuracy and scalability of ML-based ISA identification methods. Two methods are used to enrich voice analysis for depression detection: graph transformation of voice signals, and natural language processing of the transcript based on representational learning, fused together to produce final class labels ([Ghadiri et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2205.07006)). ([Uhlig et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2208.11367)) propose deep learning approximate matching (DLAM), which achieves much higher accuracy in detecting anomalies in fuzzy hashes than conventional approaches. ([Chen et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2208.14221)) present AVMiner, an expansible malware tagging system that can mine the most vital tokens from AV labels. In a correlation analysis ([Weyers et. al., 2022](https://www.paperdigest.org/paper/?paper_id=pubmed-36084447)) find no relation between intensity and dependency processing. ([Maniriho et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2209.03547)) present MalBehavD-V1, a new behavioural dataset of Windows Application Programming Interface (API) calls extracted from benign and malware executable files using the dynamic analysis approach. There is no related work to test deep learning-based malware detection models specifically. Therefore, to fill this gap ([Liu et. al., 2022](https://www.paperdigest.org/paper/?paper_id=pubmed-36107957)) propose MalFuzz.

# natural language processing in dynamic malware detection

## Anytime

([Shibahara et. al., 2016](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_glocom.2016.7841778)) propose a method for determining whether dynamic analysis should be suspended based on network behavior to collect malware communications efficiently and exhaustively. ([Nagano et. al., 2017](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1145_3022227.3022306)) propose a new method which automatically detects new malware subspecies by static analysis of execution files and machine learning. ([Karbab et. al., 2017](https://www.paperdigest.org/paper/?paper_id=arxiv-1702.05699)) address the following question: could ([Karbab et. al., 2017](https://www.paperdigest.org/paper/?paper_id=arxiv-1702.05699)) generate effective fingerprints for Android malware through dynamic analysis? The approach assessed is a novel dynamic malware analysis method, which may generalize better than static analysis to newer variants ([Kim, 2018](https://www.paperdigest.org/paper/?paper_id=arxiv-1802.05412)). ([Karbab et. al., 2019](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1016_j.diin.2019.01.017)) propose, MalDy (mal die), a portable (plug and play) malware detection and family threat attribution framework using supervised machine learning techniques. ([Qin et. al., 2020](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_icbaie49996.2020.00041)) propose a Dynamic Ransomware Detector based on the improved TextCNN(DRDT). ([Zago et. al., 2020](https://www.paperdigest.org/paper/?paper_id=pubmed-32215308)) selected 50 among the most notorious malware variants to be as exhaustive as possible. MDTA is the best suitable and manageable approach for analyzing behavioral reports using a machine learning algorithm for providing security measures to identify malware without the intervention of the investigator ([Vanjire et. al., 2021](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1007_978-981-16-1866-6_10)). ([Poudyal et. al., 2021](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_access.2021.3109260)) propose a deep inspection approach for multi-level profiling of crypto-ransomware, which captures the distinct features at Dynamic link library, function call, and assembly levels. Other influential work includes ([Mimura et. al., 2021](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1007_s10207-021-00553-8)).

## Past 5 years

The approach assessed is a novel dynamic malware analysis method, which may generalize better than static analysis to newer variants ([Kim, 2018](https://www.paperdigest.org/paper/?paper_id=arxiv-1802.05412)). ([Karbab et. al., 2018](https://www.paperdigest.org/paper/?paper_id=arxiv-1812.10327)) propose, MalDy (mal~die), a portable (plug and play) malware detection and family threat attribution framework using supervised machine learning techniques. ([Karbab et. al., 2019](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1016_j.diin.2019.01.017)) propose, MalDy (mal die), a portable (plug and play) malware detection and family threat attribution framework using supervised machine learning techniques. ([Qin et. al., 2020](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_icbaie49996.2020.00041)) propose a Dynamic Ransomware Detector based on the improved TextCNN(DRDT). MDTA is the best suitable and manageable approach for analyzing behavioral reports using a machine learning algorithm for providing security measures to identify malware without the intervention of the investigator ([Vanjire et. al., 2021](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1007_978-981-16-1866-6_10)). ([Poudyal et. al., 2021](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_access.2021.3109260)) propose a deep inspection approach for multi-level profiling of crypto-ransomware, which captures the distinct features at Dynamic link library, function call, and assembly levels. ([Yang et. al., 2021](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1186_s42400-021-00079-5)) propose a novel adversarial instruction learning technique, DeepMal, based on an adversarial instruction learning approach for static malware detection. ([Xu et. al., 2021](https://www.paperdigest.org/paper/?paper_id=arxiv-2112.10035)) present hybrid-Flacon, a hybrid pattern Android malware detection and categorization framework. ([Maniriho et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2209.03547)) present MalBehavD-V1, a new behavioural dataset of Windows Application Programming Interface (API) calls extracted from benign and malware executable files using the dynamic analysis approach. Other influential work includes ([Mimura et. al., 2021](https://www.paperdigest.org/paper/?paper_id=doi.org_10.1007_s10207-021-00553-8)).

## Past year

([Xu et. al., 2021](https://www.paperdigest.org/paper/?paper_id=arxiv-2112.10035)) present hybrid-Flacon, a hybrid pattern Android malware detection and categorization framework. ([Xu, 2021](https://www.paperdigest.org/paper/?paper_id=arxiv-2112.10038)) explore an ensemble mechanism, which presents how the combination of byte-code and native-code analysis of Android applications can be efficiently used to cope with the advanced sophistication of Android malware. To this end ([Beliaev et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2201.09369)) propose a novel defense method that consists of "truncation" and "adversarial training". ([Sahabandu et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2204.06624)) propose a binary code feature extraction model to improve the accuracy and scalability of ML-based ISA identification methods. ([Chen et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2208.14221)) present AVMiner, an expansible malware tagging system that can mine the most vital tokens from AV labels. ([Maniriho et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2209.03547)) present MalBehavD-V1, a new behavioural dataset of Windows Application Programming Interface (API) calls extracted from benign and malware executable files using the dynamic analysis approach. ([Jallad, 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2209.14727)) propose a new approach for embedding packets based on character-level embeddings, inspired by FastText success on text data. ([Abdine et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2210.05422)) propose a novel unsupervised method based on hierarchical clustering and invariant information clustering (IIC). ([Garber et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2211.06846)) use unsupervised methods in the work in order to develop a conversational pattern mining technique which does not require time consuming, knowledge demanding and resource intensive labelling exercises. Based on early-fusion and self-attention-based multimodal interaction between text and acoustic modalities ([Ghosh et. al., 2022](https://www.paperdigest.org/paper/?paper_id=arxiv-2211.14700)) propose a novel multimodal architecture for disfluency detection from individual utterances.