Elina Peronius

MACHINE LEARNING IN INTRUSION DETECTION: TOPICS FROM SCIENTIFIC LITERATURE

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2020

ABSTRACT

Peronius, Elina
Machine learning in intrusion detection: Topics from scientific literature
Jyväskylä: University of Jyväskylä, 2020, 53 p.
Cybersecurity, Master's Thesis
Supervisor(s): Lehto, Martti

Due to the traits of machine learning, many of its techniques are used in intrusion detection. Current literature of machine learning in intrusion detection lacks a good overview of the current research landscape. Due to the amount of existing data, using traditional methods to make sense of the literature would be laborious and ineffective. This study approaches the problem through using automated text analysis method called dynamic topic modelling. Dynamic topic modelling has the ability to capture the evolution of topics, which makes it a good modelling option to use on a document collection reflecting evolving content. Using the model, 21 topics were acquired, where 15 of them were deemed interpretable. Interpretable topics were labelled, though the labelling only reflects the opinion of one person. The main contribution of this study is the mapping of current research landscape. Used machine learning techniques is a well-studied area, which makes the identification of different contexts where machine learning techniques are applied in the more interesting part of the findings. Several limitations can be identified in data collection, data preprocessing, model evaluation and topic interpretation. This means that the validity of the results needs to be questioned to a degree. Due to the nature of the selected text analysis method, the results lack the richness often affiliated with traditional research methods. Due to this, suggestions of further research present topics which aim to combat this short falling. For this area of research, understanding of future evolution of topics and the identification of emerging topics would also be valuable.

Keywords: machine learning, intrusion detection, topic modelling

TIIVISTELMÄ

Peronius, Elina
Koneoppiminen hyökkäysten havaitsemisessa: Aihealueita tieteellisestä kirjallisuudesta.
Jyväskylä: Jyväskylän yliopisto, 2020, 53 p.
Kyberturvallisuus, pro gradu -tutkielma
Ohjaaja: Lehto, Martti

Koneoppimisen ominaisuudet ovat tehneet monista sen menetelmistä käytettyjä hyökkäysten havaitsemisessa. Nykyinen kirjallisuus, joka käsittelee koneoppimista hyökkäysten havaitsemissa, on vailla hyvää yleiskatsausta koko aihealueen kirjallisuuteen. Olemassa olevan datamäärän vuoksi perinteisten metodien käyttö data analyysissä olisi työlästä ja tehotonta. Tämä tutkimus lähestyy haastetta käyttämällä automaattista tekstianalyysimenetelmää nimeltä dynaaminen aihemallinnus. Dynaaminen aihemallinnus kykenee tunnistamaan aiheiden kehittymisen ajan myötä, mikä tekee siitä hyvän mallinnusvaihtoehdon käytettäväksi dokumentteihin, jotka kuvaava kehittyvää sisältöä. Dynaamisella aihemallinnuksella löydettiin 21 aihetta, joista 15 oli tulkittavia. Tulkittavat aiheet nimettiin, tosin nimeämisessä heijastuu vain yhden henkilön mielipide. Tämän tutkimuksen tärkeimmät tuotokset ovat nykyisen kirjallisuuden kartoitus. Käytetyt koneoppimisen metodit ovat hyvin tutkittu alue, joka tekee niiden kontekstien, joissa näitä menetelmiä käytetään tunnistamisesta mielenkiintoisemman osan löydöksistä. Useita puutteita tunnistettiin datan keräyksessä, datan prosessoinnissa, mallin evaluoinnissa ja aiheiden tulkinnassa. Tämän vuoksi tulosten validiteetti pitää joissain määrin kyseenalaistaa. Valitun tekstianalyysimenetelmän ominaisuuksien vuoksi tuloksista puuttuu rikkaus, joka yleensä liitetään perinteisiin tutkimusmenetelmiin. Tämän vuoksi lisätutkimuksien aiheiksi ehdotetaan aiheita, jotka pyrkivät korjaamaan tämän puutoksen. Tälle aihealueelle löydettyjen aiheiden tulevaisuuden kehittyminen ja uusien aiheiden ilmaantumisen tunnistaminen olisivat myös hyödyllisiä.

Avainsanat: koneoppiminen, hyökkäysten havaitseminen, aihemallinnus

KUVIOT

TAULUKOT

SISÄLLYS

1 INTRODUCTION

Many developments in computing and connectivity have been experienced throughout the years (Alpaydin, 2016; Gollmann, 2011). This has left assets exposed to internet facing a threat (Ghosh, Wanken, & Charron, 1998). Intrusion detection attempts to detect actions that would compromise the assets (Patcha & Park, 2007; Yu & Tsai, 2011) by monitoring the events and analysing the actions for signs of intrusions (Denning, 1987; Ghosh et al., 1998; Kemmerer & Vigna, 2002; M. Esmaili, B. Balachandran, R. Safavi-Naini, & J. Pieprzyk, 1996; Mukherjee, Heberlein, & Levitt, 1994). As more and more data needed for effective intrusion detection is collected, new methods of detection have been utilised. One of such a method is machine learning, where many of the security related problems can be approaches using learning algorithms (Yu & Tsai, 2011).

Use of machine learning has been readily adopted in intrusion detection research, a fact which can be seen on the amount of studies dedicated to the topic. This literature in turn has been studied to gain insight. When it comes to gaining insight from literature, the studied topics are pre-defined and limited to a single point of view of the overall literature. To best of knowledge, there are no studies attempting to give an overview of the research area as a whole.

Considering the amount of research dedicated to this area, a clear gap in knowledge can be identified. Considering the dependency on electronic information processing, communication networks and infrastructure that supports them (Sisäministeriö, 2017), guaranteeing cyber security of them is crucial. There are many ways to strengthen cybersecurity, such as providing new insight, as is the goal of this study.

The motivation of this study is thus two-fold. First, to offer insight in order to help strengthening cyber security and second, to fill the current gap in research. Also, as the area is continuously evolving due to changes in environments, the collected information needs to be updated.

The aim of this study is to give an idea of what intrusion detection and machine learning are, and then attempt to identify overreaching areas of interest currently present in literature. This study approaches the problem by per-

forming an empirical mono method qualitative study. The research attempts to answer following questions:

1. What overreaching areas of interest are present when considering machine learning together with intrusion detection?
2. What topics can be found when considering intrusion detection and machine learning together?
3. How do the topics evolve over time?

Used data consists of 2717 documents consisting of scientific literature collected using Scopus database. The data is analysed using automated text analysis method called dynamic topic modelling. Though the selected analysis method proved to be able to answer the research questions, the information gained was shallow. It can be argued that information gained from this study can be used by the research community. Mainly many areas of interest were identified, which for example tells what contexts are being considered. Similarly, it also tells what areas are not currently considered and thus where research is needed. This type of information can also be used by developers, as they can use this information when attempting to develop new solutions. However, not much additional contextual information can be identified through this method. Also, many limitations were found, which lowers the validity of the results.

The content is organized as follows: First, the theory base is presented in the form of exploring the main topics: intrusion detection and machine learning. In this chapter the two main topics – intrusion detection and machine learning – are presented to the reader. For intrusion detection, two of its subtypes - anomaly detection and misuse detection - are selected to be elaborated on. As for machine learning, the different learning types – supervised, unsupervised and reinforcement - are elaborated on further. After this, previous studies attempting to study the existing literature are presented. In chapter 4 research questions are presented, which is followed by the empirical part of the study. Here the selected methods are explained, and the overall process of the study is explored. In the chapter 6 (Analysis), results gained are presented and research questions answered. For the ease of understanding and followability, each research question is represented as a sub-chapter. Following the research results, findings are discussed, and conclusions drawn.

2 THEORY

In this chapter the two main topics – intrusion detection and machine learning – are presented to the reader. For intrusion detection, two of its subtypes - anomaly detection and misuse detection - are selected to be elaborated on. As for machine learning, the different learning types – supervised, unsupervised and reinforcement - are elaborated on further.

2.1 Intrusion detection

Many developments in computing and connectivity have been experienced. Computers used to be too expensive to be acquired by individuals, being affordable only to large organizations. However, as computers became cheaper, they also became available to a larger selection of the population. (Alpaydin, 2016.) As with connectivity, in 1990s instead of being isolated or connected to a local area network, computers were being connected to the internet (Alpaydin, 2016; Gollmann, 2011). This has ramifications, as the system owner no longer controls who can sent inputs to the computer or what input is sent (Gollmann, 2011). This in turn means a threat to assets exposed to the internet (Ghosh et al., 1998). In present time the expanding use of internet has left almost all sectors of society dependent on electronic information processing, communication networks and infrastructure that supports them. This dependency means a grave threat to different parts of society and its vital information systems (Sisäministeriö, 2017.)

Intrusion prevention techniques are used to protect computer systems from the threats. However, prevention alone is not enough because of the systems' complexity. (Yu & Tsai, 2011.) Intrusion detection attempts to detect actions that attempt to compromise the system or network (Patcha & Park, 2007; Yu & Tsai, 2011). This is done by monitoring the events in a system or network and analysing the actions for signs of intrusions. This happens either in real

time or after the fact. (Denning, 1987; Ghosh et al., 1998; Kemmerer & Vigna, 2002; M. Esmaili et al., 1996; Mukherjee et al., 1994.)

Intrusion can be defined to be the inappropriate access or usage of a computer or the resources of a computer system. These violations can be initiated either by outsiders attempting to break into a system or by insiders attempting to misuse their privileges. (M. Esmaili et al., 1996; S. E. Smaha, 1988; Yu & Tsai, 2011.) A successful intrusion would cause loss of one or more elements of the CIA triangle, which are confidentiality, integrity and availability (Gollmann, 2011; Mukherjee et al., 1994; Yu & Tsai, 2011). Confidentiality means that only authorized personnel should have access to the information. Integrity assures that information is accurate and that unauthorized modification doesn't happen. Availability guarantees that authorized people can access the information or resources. (Gollmann, 2011.)

Detection of attacks in the late 70's and early 80's used audit logs, in the early 90's, real-time intrusion detection systems were developed enabling detection of attacks as they occurred. After these developments, effort has been put in developing products that can be effectively deployed in large networks. (Kemmerer & Vigna, 2002.)

Accurate intrusion detection demands reliable and complete data about the activities being analysed (Kemmerer & Vigna, 2002). Data sources for intrusion detection include auditing done by an operating system, which provides operations logs. These logs might be limited to security-relevant events or they might cover every single system call invoked by every process. Another form of data collection is done by routers and firewalls, which provide event logs for network activity. (Kemmerer & Vigna, 2002; U. Lindqvist & P. A. Porras, 1999.) Though this data can be analysed in real time (U. Lindqvist & P. A. Porras, 1999) it is usually stored either indefinitely for later reference or temporarily awaiting processing (Axelsson, 1998; U. Lindqvist & P. A. Porras, 1999). This collected data is from which the intrusion detection decisions will be made from. One or more algorithms are executed to find evidence of suspicious behaviour in the audit trail. (Axelsson, 1998.)

Four different alarms or results can be generated: False alarms are classified as either being false positive or false negative. A false positive occurs when an event is reported as an intrusion, when it is in fact a legitimate activity. False negative describes the failure to detect an attack. True negative and true positive describe the successful detection of legitimate events or intrusions. (Patcha & Park, 2007.)

After the detection of possible attacks, there comes the response to said attacks, which can take many forms, such as generating an alert describing the detected intrusion (Axelsson, 1998; Ghosh et al., 1998; Kemmerer & Vigna, 2002). Generating an alarm is an example of a passive reaction. Another form of reaction is active, which takes corrective or proactive actions (Debar, Dacier, & Wespi, 1999).

Intrusion detection techniques can typically be categorised into two main approaches: anomaly detection and misuse detection (Axelsson, 1998; Debar et

al., 1999; Ghosh et al., 1998; Kemmerer & Vigna, 2002; Mukherjee et al., 1994). However, since both of the methods have their strengths and weaknesses, as will be explained in the following sub-chapters, several intrusion detection approaches should be combined to address the various intrusion threats (Lunt, 1993).

2.1.1 Anomaly detection

Earliest work on anomaly detection was introduced by James Anderson in 1980, when he presented the idea that a group of attackers identified as masqueraders could be detected by monitoring a systems audit trail for user activity that deviated from established patterns of usage (Anderson, 1980). From this early work, a more precise description of the basic assumption of anomaly detection can be drawn. The assumption is that attacks differ from normal behaviour exhibited by either a user or an application. Using this knowledge, looking for behaviour that is not typical will result in finding possible offenders. (Denning, 1987; Ghosh et al., 1998; Kemmerer & Vigna, 2002; Lunt, 1993; S. E. Smaha, 1988.)

An anomaly detection approach requires the definition of the norm. This happens through creating a model of normal behaviour of the object being monitored, be it user, system, network or program activity. The model should be created under normal operational conditions. (Ghosh et al., 1998; Patcha & Park, 2007; S. E. Smaha, 1988.) A user model may be based on information about the individual users' past behaviour, or based on generic notions of acceptable behaviour for a group of users (S. E. Smaha, 1988). After training, the learned profile is applied to new data and any activity that deviates from the baseline is treated as a possible intrusion (Patcha & Park, 2007).

The main advantage of anomaly detection is often stated to be the ability to detect previously unknown attacks (Ghosh et al., 1998; Kemmerer & Vigna, 2002; Patcha & Park, 2007). Anomaly detection can also detect variants of known attacks, and deviations from normal usage of programs regardless the user type generating them. (Ghosh et al., 1998). Since the system is based on customized profiles, it's also very difficult for an attacker to know what activity can be performed without setting of alarms (Patcha & Park, 2007).

On the other hand, the need for a training is a drawback, since the security system can't be taken into use immediately. The profiles also need upkeep, which is time-consuming. (Patcha & Park, 2007.) However, main drawback is generally stated to be the high false alarm rates (Debar et al., 1999). Both high false positive and high false negative rates can be resulted in depending on the way the algorithm is trained (Ghosh et al., 1998). Since only anomalous events are looked for, even well-known attacks can go undetected (Ghosh et al., 1998; Ilgun, Kemmerer, & Porras, 1995). This tells of the difficulties of creating a profile of normal behaviour. Entire scope of the behaviour of the target may not be covered during the learning phase. Also, behaviour can change over time, which means that behaviour profiles need to be periodically retrained. (Debar et al., 1999.) If a malicious user knows they are being modelled, they can change

their profile over time. This means that detection algorithm can be trained to treat malicious behaviour as normal. (Ghosh et al., 1998; Patcha & Park, 2007; S. E. Smaha, 1988.) Introducing new users and objects into the target system also raises problems as there is no profile information on the new users' behaviour. In addition, new users may be inexperienced with the system, which would lead to many anomaly records. (Denning, 1987.)

2.1.2 Misuse detection

Where anomaly detection benefits in finding novel attacks through behaviour profiles, misuse detection as a technique relies on pre-defined attack signatures (Patcha & Park, 2007; S. E. Smaha, 1988). Intrusions are detected by matching the pre-defined attack signatures with the collected audit trails (Debar et al., 1999; Ko, Ruschitzka, & Levitt, 1997; Patcha & Park, 2007; S. E. Smaha, 1988). When a matching signature is detected, an alarm is triggered (Debar et al., 1999). Misuse detection is popularly used, and intrusion detection systems primarily rely on it (Ghosh et al., 1998; Kemmerer & Vigna, 2002).

Misuse detection can guarantee the detection of an intrusion if a signature of the intrusion is known and included in the system (Ko et al., 1997). This detection of known attacks is also done with a low false alarm rate (Debar et al., 1999; Kemmerer & Vigna, 2002; Patcha & Park, 2007). Also, the existence of specific attack sequences ensures that it is easy for the system administrator to determine exactly which attacks are experienced. Another benefit is that the signature detection system begins protecting the system immediately upon installation. (Patcha & Park, 2007.)

The clear downside of this detection method is that if the attack signature isn't known, no alarms are raised (Patcha & Park, 2007). This means novel attacks cannot be detected, and simple variations of common attacks can go undetected (Ghosh et al., 1998; Kemmerer & Vigna, 2002; Ko et al., 1997; S. E. Smaha, 1988). To keep these systems effective, signature databases need to be maintained and updated periodically, which is a time-consuming task (Debar et al., 1999; Kemmerer & Vigna, 2002; Patcha & Park, 2007).

2.2 Machine learning

As rise of personal computers and parallel connectivity of them is experienced more and more data is created (Alpaydin, 2016). Any meaningful information is being buried in data archives too large and complex to make sense by humans (Shalev-Shwartz & Ben-David, 2014). This problem has given way to solutions which can analyse and extract information automatically and which exhibit what can be described as learning. One of such a solution is machine learning. (Alpaydin, 2016.)

Learning is the process of converting experience into expertise or knowledge (Portugal, Alencar, & Cowan, 2018; Shalev-Shwartz & Ben-David, 2014). Humans learn from experience because of the ability to reason. Computers don't have the ability to reason, and learning happens through algorithms. (Portugal et al., 2018.) Machine learning can thus be defined as the ability of a program to learn and improve their performance on a task over time through experience without the need to be explicitly programmed (Alpaydin, 2016; Patcha & Park, 2007; Shalev-Shwartz & Ben-David, 2014).

For many problems it is easier to use machine learning rather than to program a solution manually (Jordan & Mitchell, 2015). There are aspects of a given problem which may call for the use of machine learning. Tasks performed by humans, such as driving, speech recognition and image recognition, often call for the use of machine learning. Another type of task is those problems beyond human capabilities, whereas the third type of task has to do with adaptivity. (Shalev-Shwartz & Ben-David, 2014.)

The field of machine learning is sufficiently young, and even though it has progressed dramatically, it is still expanding (Jordan & Mitchell, 2015). It has also been adopted by many different fields, including intrusion detection, where security related problems can be formulated as learning problems (Yu & Tsai, 2011). In the context of intrusion detection, machine learning is used to answer to the issues of high false alarm rates. Machine learning also answers to the need of adaption to changing malicious behaviour. (Zamani, 2013.)

Learning is a very wide domain. This has led to the field of machine learning to branch into several subfields each dealing with different type of learning tasks. They can be classified based on the approach used for the learning process. Main subfields include supervised, unsupervised, and reinforcement learning. (Portugal et al., 2018; Shalev-Shwartz & Ben-David, 2014.) On top of these three main types, there are blends such as semi-supervised learning and discriminative training (Jordan & Mitchell, 2015).

2.2.1 Supervised learning

Supervised learning algorithms attempt to map an input to a correct output. This happens using training data, which includes pairs of inputs and corresponding correct outputs. (Alpaydin, 2016; Shalev-Shwartz & Ben-David, 2014.)

After training, a test set is used to measure the models prediction accuracy, which is also one of the main criteria in accepting the trained model (Alpaydin, 2016). The goal of the training is not simply to remember the training data correctly, but to learn a general model usable beyond the training examples (Alpaydin, 2016; Domingos, 2012). This means that the training data needs to reflect the characteristics of the underlying task (Alpaydin, 2016).

Though additional information provided by the supervisor is useful, it can be a source of bias and impose artificial boundaries. There is also the risk of error in labelling. (Alpaydin, 2016.)

In supervised learning, there are two different learning problems: categorical and regression. Regression is talked of when the output is numerical and classification when the output is categorical (Aggarwal & Yu, 1999; Witten & Hall, 2011).

Supervised learning includes the use of techniques such as decision tree, k-nearest neighbour, neural networks and Bayesian classifiers (Aggarwal & Yu, 1999). For intrusion detection, some techniques used are support vector machine, neural networks, decision trees and k-nearest neighbour (Androcec & Vrcek, 2018; Apruzzese, Colajanni, Ferretti, Guido, & Marchetti, 2018).

With support vector machine, the goal is to find the optimal hyper plane, or line, which separates the data into two categories. Data points, which are near the hyperplane are called support vectors. Support vectors are used to maximize the margin between them and the hyper plane. (Ayodele, 2010.)

Decision trees are structures that classify instances by sorting them based on feature values. A decision tree consists of nodes, leaves and branches. A node specifies an attribute by which the data is to be partitioned, each branch represents a possible outcome of the attribute and each leaf represents a class label. The sorting starts from the root node, which is the representation of the entire data set. When going down, the data set gets divided into smaller sets. (Kotsiantis, 2007; Sinclair, Pierce, & Matzner, 1999.)

2.2.2 Unsupervised learning

With unsupervised learning the aim is to process the input data to find the hidden structure in the data. This structure could be patterns which occur more often than others. With unsupervised learning, there is no predefined output or right answers as is in supervised learning. (Alpaydin, 2016.) This also leads to there being no distinction between training and testing data (Shalev-Shwartz & Ben-David, 2014). Unlike with supervised learning, where test set can be used to measure prediction accuracy, with unsupervised learning there is no direct measure of success (Yu & Tsai, 2011). Despite the issue of not having a direct measure of success, unsupervised learning is an important research area because unlabelled data is a lot easier and cheaper to find than labelled data (Alpaydin, 2016).

One method for unsupervised learning is clustering, where the aim is to discover the natural groupings of a set of unlabelled items (Alpaydin, 2016; Jain, Murty, & Flynn, 1999; Jain, 2010; Witten & Hall, 2011). Such a grouping also allows identifying outliers (Alpaydin, 2016). Items in the same group are similar to each other, while being different to the items in a different cluster (Jain et al., 1999; Jain, 2010). Clusters however can differ in terms of their shape, size and density (Jain, 2010). The success of clustering is often measured in terms of perceived usefulness to the users (Witten & Hall, 2011). However, the interpretation of clusters requires domain knowledge (Jain, 2010). Association rules are about discovering interesting relationships between items in database (Aggarwal & Yu, 1999).

Most frequently utilized methods under unsupervised learning include association rules, cluster analysis, self-organizing maps, and principal component analysis (Yu & Tsai, 2011).

2.2.3 Reinforcement learning

Reinforcement learning is the third of the explored learning types. It is different from the previously discussed methods in several ways. Unlike with the other learning algorithms, with reinforcement learning, there is no external source to provide the training data. The decision maker generates data by trying out different actions and receiving feedback or rewards. The decision maker then uses the feedback to update its knowledge. In time, the decision maker learns to perform the actions yielding the highest reward. Usually the decision maker has multiple possible actions to choose from and the solution to a problem often requires multiple actions. (Alpaydin, 2016.)

3 PREVIOUS RESEARCH

In this chapter studies which have attempted to gain insight or information from scientific literature are pointed out. From this literature, it is easy to see that much attention has been put in understanding the use of machine learning techniques and their strengths and weaknesses. Tsai, Hsu, Lin & Lin (2009) performed a review of used techniques in the period between 2000 and 2007. They found K-nearest neighbour and support vector machine to be the most commonly used techniques of single approach on intrusion detection. For hybrid classifiers, integrated-based hybrid classifies were the most considered. (Tsai, Hsu, Lin, & Lin, 2009.) Shashank and Balachandra (2018) made comparisons between various machine learning techniques using KDD'99 intrusion detection dataset (Shashank & Balachandra, 2018).

Li, Qu, Chao, Shum, Ho & Yang (2018) reviewed the existing intrusion detection techniques and employed KDD99 dataset for the evaluation of the machine learning-based network intrusion detection systems. Their results showed that all the approaches achieved a high detection performance in the normal, denial of service, and probes category. Conventional artificial neural network-based network intrusion detection systems led to an extremely poor detection performance in the case of user-to-root attacks and remote-to-user attacks. (Li et al., 2019.)

Apruzzese, Colajanni, Ferretti, Guido & Marchetti (2018) presented an analysis of machine learning techniques applied to the detection of intrusion, malware, and spam. Their results provide evidence of the several shortcomings that still affect machine learning techniques. All approaches were found to be vulnerable to adversarial attacks and require continuous re-training and careful parameter tuning. Moreover, when the same classifier is applied to identify different threats, the detection performance is very low. (Apruzzese et al., 2018.) Mishra, Varadharajan, Tupakula & Pilli (2019) arrived at the same conclusion when performing an analysis on various machine learning techniques and comparing them in terms of detection capability. The analysis reveals that if a technique performs well on detecting an attack, it may not perform the same for detecting other attacks. (Mishra, Varadharajan, Tupakula, & Pilli, 2019.)

Phadke, Kulkarni, Bhawalkar & Bhattad (2019) provide a survey of the proposed machine learning based intrusion detection systems (Phadke, Kulkarni, Bhawalkar, & Bhattad, 2019). Chatopadhyay & Manojit (2018) attempted to examine the progress of research in intrusion detection, which were based on machine learning techniques. They discuss the most popular machine learning techniques and their advantages and disadvantages. Most popular techniques for intrusion detection were found to be genetic algorithm, perceptron, support vector machine and fuzzy logic. (Chattopadhyay, Sen, & Gupta, 2018.)

A popular sub-area seems to be internet of things, where the use and effectiveness of techniques is analysed from a more specific point of view. Androcec & Vrcek (2018) selected and analysed 26 studies to classify the research on machine learning for the internet of things security. Most mentioned machine learning algorithms were found to be support vector machine, artificial neural network, naïve Bayes, decision tree, K-nearest neighbour, k-means clustering, random forest and deep learning. (Androcec & Vrcek, 2018.) Tabassum, Erbad & Guizani (2019) classified and categorized the intrusion detection approaches for internet of things networks, with more focus on hybrid and intelligent techniques (Tabassum, Erbad, & Guizani, 2019).

Zolanvari, Teixeira, Gupta, Khan & jain (2019) performed a literature review of the available intrusion detection solutions using machine learning models. They also deployed backdoor, command injection, and structured query language injection attacks against the system and demonstrated how machine learning based anomaly detection system performs in detecting these attacks. (Zolanvari, Teixeira, Gupta, Khan, & Jain, 2019.)

From this chapter, it can be concluded that a lot of interest is put towards understanding the used machine learning techniques, their advantages and disadvantages. A popular sub-area of interest is found to be internet of things. The studies are either based on pre-selected topics, as is with internet of things or offer a very general point of view. Those studies which offer a general point of view do not consider in what context machine learning techniques are used in. Neither option gives a good overview of the current research landscape. However, they can be used to get a good picture of the pre-selected topics.

4 RESEARCH QUESTIONS

This study attempts to study the use of machine learning in intrusion detection. It was found that prior studies interested in gaining insight are either focused on pre-selected topics or approach the issue from a general point of view, with little interest in the context the techniques are used in. Neither option can be used to give a good overview of the current research landscape. This study aims to answer to this lack by exploring the literature to find areas of interest from it. The research questions are derived from the stated aim. The questions consist of a main question and two sub questions, which will be used in answering to the main question.

> RQ1: What overreaching areas of interest are present when considering machine learning together with intrusion detection?

There are, to best of knowledge, no prior studies which would give a full overview of use of machine learning in intrusion detection. Only some sub-areas, like internet of things, are well covered. Research question 1 aims to gain more insight on the selected research area. This means that no specific sub-area is being selected beforehand. Rather the whole point is to discover possible sub-areas.

> RQ2: What topics can be found when considering intrusion detection and machine learning together?

For research question 2 the thought is that literature holds specific discoverable topics, such as different machine learning techniques or contexts they are used in, such as internet of things. Via research question 2 these topics are attempted to be identified.

> RQ3: How do the topics evolve over time?

Machine learning and intrusion detection are both continuously changing areas. The research area is not the same today as it was, say 10 years ago. New concepts have emerged while others have fallen from interest. Through mapping of the topic evolution, more information of the topics can be gained.

5 RESEARCH DATA AND METHODOLOGY

In this chapter, selected research methodology is presented. This chapter also covers selected research methods and description of the research process. Research process has been divided into six different steps. First five steps will be covered in this chapter. The final step, which is result interpretation, will be done in the following chapter.

5.1 Methodology

The term methodology refers to the way in which problems are approached (Taylor, Bogdan, & DeVault, 2015) and the selection of a methodology should be based on the purpose of the study (Taylor et al., 2015). Selected research methodology can take the form of a quantitative, qualitative or mixed methods (Saunders, Lewis, & Thornhill, 2019).

It was found that there are no prior studies which would give a good overview of the research landscape of machine learning in intrusion detection. The aim of the study is to gain insight and to fill a gap in current research. This will be done by finding overreaching areas of interest. From this stated aim, it was concluded that following a qualitative research design would be suitable. More precisely a mono method qualitative study is performed, as only a single data collection and data analysis method is used (Saunders et al., 2019).

The research follows a 6-step process, which covers the used methods for data collection and analysis:

1) Data selection and collection
2) Data pre-processing
3) Dynamic topic modelling
4) Model training
5) Evaluation of the model
6) Interpretation of results

For all studies, selecting the right data type is crucial for success. Selected data varies depending on selected research question and methodology. The research questions determine what type of data is needed and methodology what data collection methods can be used. In the first step of the process (chapter 5.2), reasoning behind data selection is given, which is followed by exploration on how the data collection was done.

Following data collection, a well thought and well executed data pre-processing is crucial since it influences the results. There are many different pre-processing tasks from where to choose from and it's clear that no single right answer to the pre-processing can be given. Therefore, there is need to understand how each task affects the data and how each task has been used before. For this prior research can be used as a basis, though it is important to note that different data types require different types of pre-processing. In the data pre-processing step (chapter 5.3), previous studies using topic modelling are analysed to find out how they have approached the problem. After this the selected tasks are presented, and the process is explained.

For the analysis method, dynamic topic modelling is selected to be used. This topic model gives us predefined number of topics and makes it possible to analyse the evolution of the terms present in said topics. In the topic modelling sub-chapter (5.4), topic models are explored to give an overview of what they are and where they can be used.

After the explanation of the analysis method, the selected model is trained. In the model training sub-chapter (5.5) the way of training is presented and any selections affecting the results are given. After the training, the model is evaluated (5.6). The last part of the process is the interpretation of the results gained from the topic modelling. This will be done in chapter 6.

5.2 Data selection and collection

In selecting suitable data for a study, there is need to make sure that the research questions can be answered through it. To find overreaching areas of interest, there is need for data that is descriptive enough. On the other hand, in order to answer the sub question on topic evolution, a long enough time period is needed to be analysed. This means that data which is cumulated over multiple years is needed. To answer these problems, scientific literature was deemed to be exceptionally good source of data, as it satisfies both criterions.

Scientific literature is available in large quantities and is easily searched and attainable through different databases, such as IEEE, Scopus and ScienceDirect. The selected data source for this study is an online research database Scopus, which is the largest abstract and citation database of peer-reviewed literature and allows the exportation of citation information to many forms like RIS, CSV and plain text.

Literature search restricted to predefined conditions was implemented to create the dataset which consists the articles' title, year of publication, abstracts and the author(s). To this end, the following search-query was implemented on the Scopus database:

((intrusion AND detection OR attack AND detection OR "intrusion detection" OR "attack detection") AND (machine AND learning OR "machine learning")) AND (EXCLUDE (PUBYEAR , 2020)) AND (LIMIT-TO (ACCESSTYPE(OA))) AND (LIMIT-TO (LANGUAGE , "English"))

With this search query, done in 23.4.2020, a dataset consisting of 2724 documents was created. Seven of the acquired documents didn't include abstract, so they were excluded from the dataset. Thus, the final number of documents is 2717. The dataset includes all publication years, expect the currently unfinished 2020 and excludes documents which are not written in English. The distribution of the documents by years is given in Table 1. From it, it can be seen that the first years of publications had very few published documents. In year 2006 double digits where reached. Thereafter number of publications started to rise almost every year.

TABLE 1 Years of publication coupled with number of documents published

Year	Count	Year2	Count2
2019	1089	2008	19
2018	533	2007	5
2017	277	2006	14
2016	242	2005	6
2015	180	2004	1
2014	113	2003	2
2013	81	2002	1

2012	57	2000	1
2011	49	1996	1
2010	22	1986	1
2009	23		

Only the abstracts from the publications are used in this study. Using full text was deemed to be unnecessary because abstract is meant to represent the original work and capture topics and key concepts in a research paper. The dataset was extracted as a comma separated value (.csv) file.

5.3 Data pre-processing

In this step of the process, a corpus is taken as an input and as an output, a pre-processed corpus is given. Data pre-processing usually contains tasks such as tokenization, stop-word removal, lowercase conversion, and stemming (Uysal & Gunal, 2014). Removing numbers is also a possible task (Karl, Wisnowski, & Rushing, 2015).

There are varying opinions when considering the importance and effect that the different pre-processing tasks have (Karl et al., 2015; Toman, Tesar, & Jezek, 2006; Uysal & Gunal, 2014). To solve this dilemma, previous set of studies that use dynamic topic modelling as an analysis method are used to guide the selection of data pre-processing tasks.

Ayele & Juell-Skielse (2020) removed stop words, punctuations, and numbers. They also removed words which were deemed irrelevant and trivial through term frequency analysis and word clouds. They used lemmatization to achieve the root forms of terms. (Ayele & Juell-Skielse, 2020.) Ha, Beijnon, Kim, Lee & Kim (2017) performed word tokenization, parts-of-speech-filtering, stop-words filtering and stemming (Ha, Beijnon, Kim, Lee, & Kim, 2017). Greene (2017) used standard case conversion, tokenization and lemmatization. They removed short tokens and token corresponding to generic stop words, parliamentary-specific stop words and names of politicians. They also removed tokens that occurred in less than 5 speeches. (Greene, 2017.) Blei and Lafferty (2006) used stemming, removed function terms and removed terms that occurred fewer than 25 times (D. Blei & Lafferty, 2006). Considering the prior, five pre-processing tasks were selected and applied to the corpus:

1) Lowercase conversion
2) Number removal
3) Tokenization and punctuation removal
4) Stopword removal

5) Lemmatization

The data pre-processing starts with lowercase conversion. This converts all uppercase characters to their lowercase forms. After this numbers were removed, as for this topic, they don't hold any meaningful information and aren't an essential part of the corpus. The effects of lowercase conversion can be seen in figure 1, where excel was used. Number removal was done using regex in python.

```
malicious domain name attacks have become a serious issue for internet security. in this study, a malicious domain names detection algorithm based on n-gram is proposed.
```

FIGURE 1 Sample piece of document after lowercase conversion

Tokenization was applied using NLTK RegexpTokenizer. In tokenization text is split into parts, here into words. Figure 2 depicts this. RegexpTokenizer also deletes punctuations, which simplifies the pre-processing as another additional step isn't needed for it.

```
['malicious', 'domain', 'name', 'attacks', 'have', 'become', 'a', 'serious', 'issue', 'for', 'internet', 'security', 'in', 'this', 'study', 'a', 'malicious', 'domain', 'names', 'detection', 'algorithm', 'based', 'on', 'n', 'gram', 'is', 'proposed']
```

FIGURE 2 Sample piece of document after tokenization

Stop-words represent words that are found commonly in any sentences and occur very frequently. These words, if not removed would be overrepresented in modelling results, without giving much meaningful information. The standard NLTK English stopword list was used to remove them from the corpus. As seen in the figure 3, terms such as "a", "for" and "is" are removed.

```
['malicious', 'domain', 'name', 'attacks', 'become', 'serious', 'issue', 'internet', 'security', 'study', 'malicious', 'domain', 'names', 'detection', 'algorithm', 'based', 'n', 'gram', 'proposed']
```

FIGURE 3 Sample piece of document after stopword removal

Previous studies have used both stemming and lemmatization to find the root form of words. There is thus need to explain why lemmatization was selected for this study. Both lemmatization and stemming are used to obtain the root form of a word and to reduce word variation. They differ on how the root word is obtained and what type of root word is gained. Essentially stemming can result in non-words, while lemmatization only produces actual words.

To find the best choice between lemmatization and stemming, two stemming packets (PorterStemmer and SnowballStemmer) and a lemmatization packet (WordNetLemmatizer) were used on tokenized corpus and their results compared. The evaluation is based on how well the correct root form is found and how interpretable the resulting words are. Interpretability will be important when analysing the results of topic modelling.

Considering how well the correct root is found, lemmatization performs the best out of the three. With the stemmers, words were often pruned so far that the correct root word was lost. Selecting stemmers would thus make it harder to explain modelling results. Figure 4 shows how lemmatization finds the root words.

```
['malicious', 'domain', 'name', 'attack', 'become', 'serious', 'issue', 'internet', 'security', 'study',
'malicious', 'domain', 'name', 'detection', 'algorithm', 'based', 'n', 'gram', 'proposed']
```

FIGURE 4 Sample piece of document after lemmatization

5.4 Topic modelling

The amount of digital data produced is growing at an increasing pace and to make sense of it, new techniques have been developed. This type of change has also reached the research community, where automated analysis methods have emerged beside traditional text analysis methods. From these automated analysis methods, inductive methods as well as supervised methods can be identified. (Purhonen & Toikka, 2016.) With automated analysis methods, work traditionally done by researcher can be automated either fully of partly (Mills, 2017). For studies that use large amounts of data, automated analysis methods are ideal, as analysis using traditional methods would be if not impossible, certainly laborious and ineffective.

For this study, an inductive analysis method called topic modelling, is used. Topic modelling is a term for a wide variety of algorithms, which aim to discover themes or topics from texts through word analysis. Topic is defined as a distribution over a fixed vocabulary. (D. M. Blei, 2012; Ignatow & Mihalcea, 2017.)

Topic modelling has been found to be of benefit to many qualitative studies (Nikolenko, Koltsov, & Koltsova, 2015). Topic modelling has also been employed on a wide variety of texts, including political texts (Purhonen & Toikka, 2016), social media feeds (Rohani, Shayaa, & Babanejaddehaki, 2016) and scientific literature (Gurcan, 2019; Gurcan & Sevik, 2020).

There are many different types of topic models, with the most prominent being latent dirichlet allocation, henceforth called LDA. LDA learns a predefined number of latent topics, where each topic is represented as a distribution over terms and each document as a distribution over topics. (D. M. Blei, Ng, & Jordan, 2003; Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009.)

Topic models cover not only the discovering of latent themes (D. M. Blei et al., 2003), but also topic change over time (D. Blei & Lafferty, 2006), and how the topics are connected to each other (D. Blei & Lafferty, 2005). Topic models have also evolved to use supervised learning instead of unsupervised learning (D. M. Blei & McAuliffe, 2007).

Considering the number of possible models from where to choose from, it is important to return to the aim of this study. The aim is to gain more insight using the selected data, which in this case is scientific literature collected from Scopus. It is important to note that certain document collections, such as scholarly journals reflect evolving content (D. Blei & Lafferty, 2006). Taking this and the fact that evolution of topics is desired into consideration, dynamic topic modelling was selected to be used.

Dynamic topic model builds on top of LDA with the ability to capture the evolution of topics in a sequentially organized corpus of documents. This is done by dividing documents by time slice. This means that while there is still no understanding of the order of words in a document, the order of documents is now accounted for. Each slice of documents is modelled with a K-component topic model. Topics and topic proportion distributions are then chained together sequentially. (D. Blei & Lafferty, 2006.) As Blei and Lafferty (2006) illustrate, dynamic topic models can capture different scientific themes, and can be used to inspect trends of word usage within them (D. Blei & Lafferty, 2006).

5.5 Training the model

There are many different software available which can perform topic modelling, such as Mallet (McCallum, 2002), Stanford topic modelling toolbox (Ramage & Rosen, 2009), Gensim (Rehurek & Sojka, 2010), and the implementation written by David Blei, published in blei Git repository. However, most of these are focused on LDA, which limits the options available.

For this study, Gensim is used. Gensim is a free Python library which offers two different ways to perform dynamic topic modelling. First one is using the wrapper for original C++ DTM code made by Blei. The second one is using a LdaSeqModel class, which is an effort to have a pure python implementation of the previously mentioned. This study uses the wrapper for the original code.

Two main inputs for the model are the corpus and the dictionary. Dictionary was created using the option to filter extremes from it. Following limitations were made: no words which are present in corpus less than 25 times and no words which are present in more than 50 percent of the documents.

After these tasks, the important decision to make is to select the number of topics. To determine the most suitable number of topics for dynamic topic modelling, gensim LdaModel was run using different options for topic number. Coherence score for each option was then counted. Topic coherency is used to tell the interpretability of the topics. Coherence value is found using Gensim CoherenceModel while using c_v as a coherence measure and setting iteration to 8. The higher the value the more coherent topics are present. This is illustrated in figure 5, where number of topics is coupled with the corresponding coherency. Topic coherency value starts low and rises as topic number rises. The peak coherency is reached around topic number 21. After this topic coherency value starts to drop. This indicates that topic number 21 should be used.

For dynamic topic modelling time slices also need to be set. Time slice is represented as the number of documents on each year. The created corpus has 21 time slices from year 1986 to 2019. First time slice holds 1 document while 21st time slice holds 1089 documents. Thus, time slices are set as so: [1, 1, 1, 1, 2, 1, 6, 14, 5, 19, 23, 22, 49, 57, 81, 113, 180, 242, 277, 533, 1089].

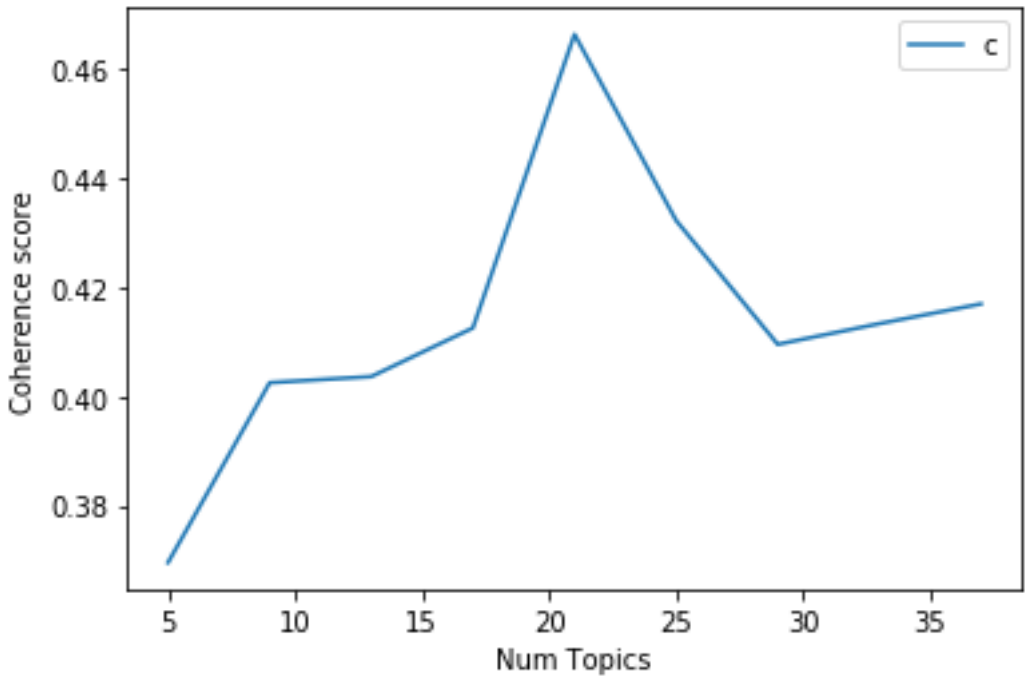

FIGURE 5 Coherence score coupled with number of topics

5.6 Model evaluation

For this study, human evaluation was used to evaluate the coherence of the created model. Evaluation was performed following example set by Xie & Xing (2013) and Zhang, Kim & Xing (2015). 10 most probable terms for each topic were picked. First the topic interpretability is judged. If interpretability is bad, the words present in this topic are labelled as "irrelevant". If the topic is deemed to be interpretable, relevant terms are identified. Similarly, to evaluation made by Zhang, Kim & Xing (2015) if more than a half of words are classified as relevant, then the topic is regarded as coherent. Where Xie & Xing (2013) used subjects to judge the topics, for this study judging is done by the researcher. After these steps, coherence measure for the model can be counted. Coherence measure is defined as the ratio between the number of relevant words and total number of words in valid topics (Xie & Xing, 2013).

As seen in table 2, 6 topics were identified as irrelevant, whereas only 9 topics were deemed coherent. Resulting coherence measure is 0,60. Considering

that the evaluation is made by one person, subjective biases are introduced to the resulting evaluation.

TABLE 2 Term relevancy, where uncoherent words are crossed out.

NO	Keywords
1	learning, data, machine, ~~model~~, deep, network, neural, training, big, ~~method~~
2	~~attack, system, security, cyber, threat, control, cloud, network, proposed, based~~
3	vehicle, ~~message~~, ~~grid~~, ~~detector~~, safety, driver, ~~time~~, bus, vehicular, road
4	detection, intrusion, network, system, id, ~~rate~~, ~~attack~~, ~~based~~, ~~proposed~~, ~~accuracy~~
5	recognition, pattern, immune, theory, ~~object~~, ~~student~~, ~~programming~~, image, ~~evidence~~, ~~multiple~~
6	iot, data, ~~research~~, application, system, device, ~~security~~, ~~paper~~, ~~technology~~, ~~computing~~
7	~~user, data, mining, rule, study, social, web, information, profile, technique~~
8	~~feature, method, algorithm, data, proposed, classification, result, based, performance, accuracy~~
9	network, traffic, packet, attack, ~~based~~, ~~flow~~, detection, protocol, node, ~~service~~
10	agent, system, authentication, ~~action~~, ~~multi~~, visual, eye, biometric, monitoring, ~~electricity~~
11	sensor, wireless, node, ~~proposed~~, ~~based~~, ~~algorithm~~, ~~mdpi, switzerland, basel, licensee~~
12	optimization, algorithm, swarm, ~~problem~~, ~~model~~, particle~~, parameter,~~ pso~~, search~~, ~~fusion~~
13	data, anomaly, detection, ~~time~~, behaviour, ~~approach~~, ~~event~~, ~~real~~, ~~stream~~, pattern
14	trust, game, strategy, ransomware, trusted, equilibrium, ~~member~~, risk, ~~phase~~, ~~study~~
15	svm, vector, machine, support, ~~signal~~, kernel, classification, ~~based~~, ~~accuracy~~, ~~feature~~
16	~~model, study, area, result, spatial, test, high, index, map, author~~
17	~~model, prediction, network, power, system, neural, energy, time, parameter, artificial~~
18	image, disease, patient, classification, medical, ~~using~~, diagnosis, classifier, region, cancer
19	~~domain, source, gene, expression, spectral, ontology, study, e, recurrent, gru~~
20	malware, ~~analysis~~, malicious, ~~method~~, detection, ~~feature~~, call, ~~code~~, android, ~~technique~~
21	fuzzy, rule, human, ~~model~~, knowledge, ~~system~~, decision, logic, cognitive, complex

6 ANALYSIS

This chapter presents the results gained from dynamic topic modelling and answers to the selected research questions. For the ease of understanding and followability, each research question is represented as a sub-chapter. As the sub-questions are meant to help in answering the main question, they are covered first in sub-chapters 6.1 and 6.2. After this knowledge gained from them is used in answering the main research question.

6.1 Topic interpretation

Research question 2 asks what topics can be found from data. Scientific literature was selected as this study's data, and corpus consisting of 2717 documents was collected. Using dynamic topic modelling, 21 topics were acquired. The topics are given as a combination of term probability and term, however in this study only the terms are used in interpretation.

The last task of topic modelling is the interpretation of the results. There is no unambiguous rule for the naming, though often the names are based on the use of topics' common or descriptive terms and the interpretation of them (Nelimarkka, 2019). It is also important to recognize topics which are either worthless or misleading (Ignatow & Mihalcea, 2017). However, the labels represent the labellers interpretation about the meaning of words and are thus subjective. In this sub-chapter, topics are explored as they are in time slice 21, which represents the year 2019. Interpretation happens based on the 10 most probable terms.

As seen in table 3, when considering terms present there is some overlapping. Through manual observation, it is obvious that some words outside stopword list are overrepresented in the corpus. Two different groups can be found. First group represents common words often found in research papers, such as "method", "proposed", "based" and "study". Second group of overrepresented words were words which would be expected to be on abstract

covering intrusion detection and machine learning such as “network”, “detection”, “security”, “system” and “attack”. For the second group, a lot of the overrepresented words are also words, which are part of a bigram or trigrams such as “neural network” or “intrusion detection system”. Overrepresented words are most likely caused by not altering stopword list to include additional words. This adding of words would however had introduced subjective biases, since the researcher would decide which word are important and which commonalities.

Important to note, some of the keywords are words, which don’t describe the abstract contents, but are a representation of a copyright string in abstracts. These were not accounted in the data pre-processing, so they make an appearance in the modelling results. These words are present in topic 11 with keywords “mdpi”, “Switzerland”, “basel” and “licensee”. These keywords together form the string “licensee mdpi, basel, Switzerland”.

6 topics were identified as hard to interpret. In addition to this, some words in the rest of the topics are not relevant to the interpretation. This aspect was covered in the model evaluation chapter (5.6). Despite the overrepresented words and some hard to explain topics, dynamic topic modelling has given unique topics, which are mostly easy enough to explain.

TABLE 3 The 21 topics as they are in year 2019 labelled based on 10 most probable words

NO	Topic Name	Keywords
1	Deep Learning	learning, data, machine, model, deep, network, neural, training, big, method
2	–	~~attack, system, security, cyber, threat, control, cloud, network, proposed, based~~
3	Vehicle	vehicle, message, grid, detector, safety, driver, time, bus, vehicular, road
4	Intrusion detection system	detection, intrusion, network, system, id, rate, attack, based, proposed, accuracy
5	Pattern recognition	recognition, pattern, immune, theory, object, student, programming, image, evidence, multiple
6	Internet of things	iot, data, research, application, system, device, security, paper, technology, computing
7	–	~~user, data, mining, rule, study, social, web, information, profile, technique~~
8	–	~~feature, method, algorithm, data, proposed, classification, result, based, performance, accuracy~~
9	Network attack detection	network, traffic, packet, attack, based, flow, detection, protocol, node, service
10	Authentication	agent, system, authentication, action, multi, visual, eye, biometric, monitoring, electricity
11	Wireless technologies	sensor, wireless, node, proposed, based, algorithm, mdpi, switzerland, basel, licensee
12	Particle swarm optimization	optimization, algorithm, swarm, problem, model, particle, parameter, pso, search, fusion
13	anomaly detection	data, anomaly, detection, time, behaviour, approach, event, real, stream, pattern

14	Game theory	trust, game, strategy, ransomware, trusted, equilibrium, member, risk, phase, study
15	Support vector machine	svm, vector, machine, support, signal, kernel, classification, based, accuracy, feature
16	–	~~model, study, area, result, spatial, test, high, index, map, author~~
17	–	~~model, prediction, network, power, system, neural, energy, time, parameter, artificial~~
18	image classification	image, disease, patient, classification, medical, using, diagnosis, classifier, region, cancer
19	–	~~domain, source, gene, expression, spectral, ontology, study, e, recurrent, gru~~
20	mobile malware detection	malware, analysis, malicious, method, detection, feature, call, code, android, technique
21	Fuzzy logic	fuzzy, rule, human, model, knowledge, system, decision, logic, cognitive, complex

From 21 topics, 6 topics are not easily interpretable. These topics are 2, 7, 8, 16, 17 and 19. The terms present in these topics don't seem to have much association with each other and no clear label can be given.

The vocabulary of the topics 1 (deep learning), 12 (particle swarm optimization), 15 (support vector machine) and 21 (fuzzy logic) consist of terms about different algorithms, including learning algorithms and optimization algorithms. These topics form a group describing the techniques mostly considered in the literature. Considering the dataset, it is not surprising that machine learning techniques, even multiple ones would be found.

The most identifying terms in topic 5 were "pattern", "recognition" and "image". Though not many clearly describing terms, a label can be set as "pattern recognition". Pattern recognition, as the name suggest concerns itself with identifying objects in a picture.

The vocabulary of topic 3 is one of the most unique, as it has no overlapping when considering terms present. It is also very intuitively interpretable to be about vehicles and driving.

Topic 6 is also intuitively interpretable. However, its vocabulary has some overlapping of terms. The vocabulary consists of terms which can often be coupled with internet of things, such as "internet of things-device", "internet of things-application" and "internet of things-system". However, these terms are also quite often used in other contexts.

Topic 4 describes intrusion detection systems. This topics vocabulary is mostly consisting of overrepresented words often found in the literature. It is explained as intrusion detection system, since it both has the abbreviation "ids" and also all the individual words, which combined form the actual multiword "intrusion detection system".

Topic 9, which was named as network protocol attacks, consist of many terms associated with network protocols. This topic is also quite easy to interpret, as it has many unique terms, which are highly associated with networks. This coupled with terms "attack" and "detection" make it easy to interpret.

Vocabulary of topic 10 in first view consist of terms, which don't create a one coherent topic. However, the presence of term authentication is the key to its interpretation. This is because many of the other terms, such as "eye", "biometric" and "agent" can be coupled with it to form a meaning of authentication issues.

Topic 11 was explained as wireless technologies due to the occurrence of terms "sensor", "wireless" and "node" which can all be associated with wireless technologies. For this topic, it is important to note that the 4 least probable words are part of copyright sting mentioned earlier.

Vocabulary of topic 13 consist of terms often associated with anomaly detection, such as "anomaly", "detection" and "behaviour". Therefore label "anomaly detection" is given. Again, with this topic only a few of the terms can be used in interpreting the topic.

Vocabulary of topic 14 consist of unique terms about strategies in games and picking the best response to an action. These terms are why it is labelled as "game theory". With this topic, there are quite many describing terms, which makes it a coherent topic.

Topic 18 was explained as image classification. It consists of terms about classification coupled with many medical terms, such as "cancer", "patient" and "diagnosis". This could point to image classification in the use of diagnosis of cancer. Topic 18 is interesting, given that it differs greatly from what would be expected from the selected literature. This topic could point to other areas of literature than the intended one being in the corpus. It is also a topic which consists of many describing terms. This makes it both easy to interpret and a coherent topic.

Vocabulary of topic 20 consist of terms about malware detection. Since the term android is also present, a more precise "mobile malware detection" label was selected.

Through this type of exploration, topic modelling paints a picture of quite many areas of interest. However, only 9 of these topics can be identified as being truly coherent with more than half terms being relevant. This means that for the most of the topics, only a few terms are used to label the topics. This also puts a lot importance on the opinion of the labeller. The results tell that machine learning techniques are considered the most in the literature. Also, different contexts were identified. There are also indications that other areas of literature than the intended one was included.

6.2 Topic evolution over time

It is argued that understanding of topic evolution can help in giving more insight on the selected study area. It is acknowledged that the studied area is evolving, which means that the topics are changing as well. The selected area and found topics are not the same in current day as they were in 1990. During this time new concepts have emerged, which should show in topic evolution through term emergence and term fluctuation.

Research question 3 asks how the topics found using dynamic topic modelling evolve over time. The perceived time here starts in 1986 and ends in 2019. This gives 21 different time slots to consider. Considering that this is quite many time slots to examine, four timeslots are sampled to further analysis. Years 1986, 2007, 2013 and 2019 are selected to be examined. As in previous sub-chapter, topics will be examined based on the 10 most probable words. By analysing the term change, topics can be understood better. However, overall the term movement was very small in all the topics and not much additional information could be had from the results. From the 21 topics, 3 topics show enough term fluctuation to analyse further. All the topic evolutions can be found in appendix 1.

6.2.1 Internet of things

This topic was interpreted as internet of things in the previous sub chapter. Considering how the term probabilities evolve throughout the timeslots, it can be seen that internet of things as term didn't clearly come up until the last timeslot considered. Of course, not all the 21 timeslots are considered, and it is probable that internet of things as a term is present earlier than 2019. What is interesting is that when considering only timeslots 1, 2 and 3, the topic is very hard to interpret.

TABLE 4 Term evolution of topic "internet of things"

1986	2007	2013	2019
system	system	system	iot
application	application	data	data
data	data	research	research
research	research	application	application
paper	paper	paper	system
information	information	information	device
security	security	security	security
technology	technology	technology	paper
provide	provide	device	technology
based	device	provide	computing

6.2.2 Wireless technologies

Topic “wireless technologies” is a topic that can be made more precise through topic evolution results. From the first 3 timeslots, the term “wsns” is present, which is the abbreviation for wireless sensor networks. Based on this information, the topic could be renamed to a more precise “wireless sensor network”. Considering that timeslot 4 includes many terms used in copyright string, it is possible that the term “wsns” would be present were the copyright terms removed.

TABLE 5 Term evolution of topic "wireless technologies"

1986	2007	2013	2019
sensor	sensor	sensor	sensor
wireless	wireless	wireless	wireless
node	node	node	node
proposed	proposed	proposed	proposed
based	based	based	based
algorithm	algorithm	algorithm	algorithm
wsns	wsns	wsns	mdpi
object	object	licensee	switzerland
licensee	licensee	object	basel
network	network	switzerland	licensee

6.2.3 Mobile malware detection

This topic was described as mobile malware detection. However, interpreting this topic in any other time slot, this wouldn’t be the description of this topic. Based on the first three timeslots, the description would have been much more general, such as malware detection. This could point to the rising interest in mobile device security.

TABLE 6 Term evolution of topic "mobile malware detection"

1986	2007	2013	2019
malware	malware	malware	malware
call	call	call	analysis
method	method	malicious	malicious
malicious	malicious	method	method
program	program	detection	detection
detection	detection	program	feature
analysis	analysis	analysis	call
source	source	source	code
code	code	code	android
use	use	use	technique

6.3 Areas of interest

Research question 1 asks what areas of interest can be found when considering machine learning and intrusion detection together. In previous sub-chapters, the 21 topics found and their evolution from year 1986 to 2019 were covered. The 14 labelled topics are identified as the areas of interest. Topic 18 (image classification) was excluded as it was explained to be more about cancer diagnosis rather than intrusion detection. The 14 topics are respectively deep learning, vehicle, intrusion detection system, pattern recognition, internet of things, network attack detection, authentication, wireless sensor networks, particle swarm optimization, anomaly detection, game theory, support vector machine, mobile malware detection and fuzzy logic. These areas of interest can be further grouped into two different categories: techniques and contexts of use.

6.3.1 Machine learning techniques

First category that can be used to group the found topics is identified as machine learning techniques. These topics are deep learning, particle swarm optimization, support vector machine and fuzzy logic. Based on the amount of the topics that can be labelled as different algorithms, this area is the most prominently considered in the literature. Considering the used data, this is not a surprise.

The topics in this group, excluding topic fuzzy logic don't offer description on the qualities of the algorithms. They don't tell whether the algorithm is considered effective, or any other qualities of them. Only exception to the rule is fuzzy logic, where the terms present describe a technique often used with terms about human cognition.

Even though multitude of techniques have been identified, no clear conclusions can be drawn on what type of learning is the most prominent one in literature. Not all the techniques are learning type specific, which means they can be considered supervised, unsupervised or reinforcement learning depending on how they are applied. Also, based on the results no associations can be drawn between them and the other found topics. The results don't tell if these topics are studies alone, or in a specific context.

6.3.2 Contexts of use

Other category that can be used to group the found topics is identified as contexts of use. These topics represent the different contexts the machine learning techniques can be used in. Some of them represent more general areas, like intrusion detection systems, network attack detection and anomaly detection, while others are more precise in their nature, like vehicles, internet of things and wireless sensor networks. Some of the topics are often closely associated with each other and could thus be grouped under a single set of headers. As an

example, internet of things and wireless sensor network could be considered together as wireless sensor networks are often used within internet of things-systems. However, this does not mean that wireless sensor networks are only studied with internet of things. Indeed, the results indicate the opposite as they were identified as separate topics. Had they been considered mostly together they would have been grouped in a same topic.

From the results the areas of internet of things and mobile malware detection describe more emergent terms than the other topics. The term "IOT" which is the most descriptive term in topic 6 only appeared in the last considered time slice. Same type of development was seen with mobile malware detection, though to a lesser degree, as the term "android" didn't emerge till time slice 4. However, the term "android" didn't have that high probability.

Identifying of different contexts is important, as they are not often considered together in the literature and identifying them manually can be ineffective. However, the results don't mean that the resulting list is complete. It is likely that additional areas of interest are present in the literature, but they are not considered consistently enough to appear in the results. Also, as with machine learning techniques, no conclusions about associations between topics can be drawn.

7 DISCUSSION

In this chapter, interpretation of what the results mean in terms of existing research in the same field is given. The limitations of the study are also presented covering all the process phases. Lastly possible future research topics are presented.

7.1 Main results

Using dynamic topic modelling 21 topics were found, with 15 being interpretable. Through the results, it was possible to answer all the research questions. This study identified 14 areas of interest, which are deep learning, vehicle, intrusion detection system, pattern recognition, internet of things, network attack detection, authentication, wireless sensor networks, particle swarm optimization, anomaly detection, game theory, support vector machine, mobile malware detection and fuzzy logic.

These areas of interest were grouped under two categories: machine learning techniques and contexts of use. Given prior literature some comparisons can be made with both categories. This study's results concur with that of the previous ones on the fact that machine learning techniques are by far the most studied area in the literature. Given that deep learning, support vector machine, particle swarm optimization and fuzzy logic were found as techniques, it can be said that the results using dynamic topic modelling doesn't give any new or contradicting information on this field. However, results lack in the area of identifying all techniques present. Previous literature has identified other techniques, such as K-nearest neighbour, naïve Bayes, genetic algorithm, decision tree and random forest (Shashank & Balachandra, 2018; Tsai et al., 2009). This can point to these techniques not holding that much importance in the literature, or the model not being able to capture them. Either way this means that the results are lacking and require additional sources of information to fill in to be comprehensive.

However, perhaps more importantly, many contexts of use were identified. This information does help fill a gap in information, when considering previous literature. Considering the found contexts, this study offers new insight by mapping the research landscape. Identifying of different contexts is important, as they are not often considered together in literature and identifying them by hand is ineffective. However, no conclusion can be made whether the resulting list of contexts is complete. It is likely that additional areas of interest are present in the literature, but they are not considered so consistently that they would appear in the results.

Even though the identification of contexts can be valuable, this is pretty much the only valuable finding this study offers. Due to the nature of the selected analysis method, a lot of valuable information which is retained using traditional qualitative analysis methods is lost. Traditional methods can recognise associations between topics and make much more insightful observations based on the associations. Prior literature identifies techniques in their contexts and draw other meaningful insight of strengths and weaknesses. Results gained with topic models however are much more generalised and lose detail.

7.2 Limitations

Considering data collection and pre-processing the quality of decisions made is hard to evaluate before the final results. Thus, found limitations are based on the end results. From the results terms, which don't seem to have any association with the intended literature can be identified. More precisely topic "image classification" was identified as such a topic. The terms present in this topic point towards medical field, more precisely to a cancer diagnosis through image classification. This shows a fault in data collection, as unintended literature has gotten through.

Moving to data pre-processing, the results show terms pointing to the copyright string. This shows a limitation in the data pre-processing phase. The fact that two different types of overrepresented words were identified also supports the notion of limitation in data pre-processing. Overrepresented words could be removed before modelling, though this introduces subjective biases. It needs to also be stated that there is no clear consensus on the effects of different data pre-processing tasks on automated analysis methods (Karl et al., 2015; Toman et al., 2006; Uysal & Gunal, 2014). This coupled with the fact that full effects and limitations of selected data pre-processing tasks can't be seen until results are produced makes selecting tasks hard.

The limitation in model evaluation is based on the fact that it was performed by a single person, which is inherently vulnerable to subjectivity. Typically human based model evaluation is performed by multiple people (Chang et al., 2009; Xie & Xing, 2013; Zhang et al., 2015) and then the results can be combined. Same limitation is present in topic interpretation, where labelling is done by the researcher. Even though there is no unambiguous rule for the naming

(Nelimarkka, 2019), it is clear that using a single person for the interpretation of the topics introduces subjective biases.

7.3 Future research

Many future areas of research can be identified. First, topic correlation should be examined. Each of the found area of interest could also be studied in a more in-depth manner, possibly through literature reviews or more precise topic modelling. Also, studying future evolution of topics and emergence of new topics could offer especially valuable information. Future studies attempting to use topic modelling should also implement a very well-done data collection and pre-processing. In data pre-processing, bi -and trigrams could be accounted for. Also, overrepresented words could be removed. As for model evaluation and topic interpretation, multiple people should be used and the need for domain knowledge should be taken into account.

8 CONCLUSION

The aim of this study was to gain insight by identifying overreaching areas of interest from the literature. Dynamic topic modelling was used to find 21 topics and coherence measure was selected as metric of the model quality. From 21 topics, 6 were deemed to be hard to interpret. The rest 15 topics were interpret using 10 most probable terms. Topic evolution was explored, and it was found that for the majority of topics there was no large term change or movement. 14 topics were identified as areas of interest and then grouped under two categories: machine learning techniques and contexts of use.

This study has contributed by offering an understanding of the state of the research literature. Most notable finding was identifying the different contexts where machine learning techniques are used in. However, due to the nature of selected analysis method, this was the extent of the notable findings. Topic model lost much context specific information, such as associations between topics.

Limitations were also identified in data collection, pre-processing, evaluation and interpretation. Results indicate that unintended literature has been included. Results also indicate limitations in data pre-processing as overrepresented words could be identified and terms indicating copyright strings found. Model evaluation and topic interpretation suffer from the same limitation, that is the use of only one person, which introduces subjective biases. Due to the limitations the validity of the findings must be questioned to a degree.

Future studies should take the limitations of this study into account to have better validity of results. Other study areas could consider topic association. Each of the found areas of interest could also be studied more in-depth. Also study of future evolution of topics and study of emerging topics would no doubt provide valuable information to research community and other interested parties alike.

REFERENCES

Aggarwal, C. C., & Yu, P. S. (1999). Data mining techniques for associations, clustering and classification. *Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data MiningAp,* , 13-23. doi:10.1007/3-540-48912-6_4

Alpaydin, E. (2016). *Machine learning : The new AI*. Cambridge: MIT Press. Retrieved from http://ebookcentral.proquest.com/lib/jyvaskyla-ebooks/detail.action?docID=4714219

Anderson, J. P. (1980). *Computer security threat monitoring and surveillance*. Fort Washington: James P. Anderson Co.

Androcec, D., & Vrcek, N. (2018). Machine learning for the internet of things security: A systematic review. *Proceedings of the 13th International Conference on Software Technologies,* , 563-570. doi:10.5220/0006841205630570

Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., & Marchetti, M. (2018). On the effectiveness of machine and deep learning for cyber security. *10th International Conference on Cyber Conflict (CyCon),* , 371-390. doi:10.23919/CYCON.2018.8405026

Axelsson, S. (1998). *Research in intrusion-detection systems: A survey*. Gothenburg, Sweden

Ayele, W. Y., & Juell-Skielse, G. (2020). Eliciting evolving topics, trends and foresight about self-driving cars using dynamic topic modeling. In K. Arai, S. Kapoor & R. Bhatia (Eds.), *Advances in information and communication* (pp. 488-509) Springer, Cham. doi:10.1007/978-3-030-39445-5_37

Ayodele, T. O. (2010). Types of machine learning algorithms. In Y. Zhang (Ed.), *New advances in machine learning* () IntechOpen. doi:10.5772/9385 Retrieved from https://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84. doi:10.1145/2133806.2133826

Blei, D. M., & McAuliffe, J. D. (2007). Supervised topic models . *Proceedings of the 20th International Conference on Neural Information Processing Systems,* , 121-128.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022. doi:10.1162/jmlr.2003.3.4-5.993

Blei, D., & Lafferty, J. (2005). Correlated topic models. *Proceedings of the 18th International Conference on Neural Information Processing Systems,* , 147–154.

Blei, D., & Lafferty, J. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning,* , 113-120. doi:10.1145/1143844.1143859

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Proceedings of the 22nd International Conference on Neural Information Processing Systems,* , 288-296.

Chattopadhyay, M., Sen, R., & Gupta, S. (2018). A comprehensive review and meta-analysis on applications of machine learning techniques in intrusion detection. *Australasian Journal of Information Systems, 22* doi:10.3127/ajis.v22i0.1667

Debar, H., Dacier, M., & Wespi, A. (1999). Towards a taxonomy of intrusion-detection systems. *Computer Networks, 31*(8), 805-822. doi:10.1016/S1389-1286(98)00017-6

Denning, D. E. (1987). An intrusion-detection model. *IEEE Transactions on Software Engineering, SE-13*(2), 222-232. doi:10.1109/TSE.1987.232894

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87. doi:10.1145/2347736.2347755

Ghosh, A. K., Wanken, J., & Charron, F. (1998). Detecting anomalous and unknown intrusions against programs. *Proceedings 14th Annual Computer Security Applications Conference (Cat. no.98EX217),* , 259-267. doi:10.1109/CSAC.1998.738646

Gollmann, D. (2011). *Computer security* (3rd ed.) John Wiley & Sons, Ltd. Retrieved from https://www.academia.edu/40748431/Dieter_Gollmann_Wiley.Computer.Security.3rd.Edition

Greene, D. (2017). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis, 25*(1), 77-94. doi:10.1017/pan.2016.7

Gurcan, F. (2019). Major research topics in big data: A literature analysis from 2013 to 2017 using probabilistic topic models. *2018 International Conference on Artificial Intelligence and Data Processing (IDAP),* doi:10.1109/IDAP.2018.8620815

Gurcan, F., & Sevik, S. (2020). Mapping the research landscape of deep learning from 2001 to 2019. *2019 1st International Informatics and Software Engineering Conference (UBMYK),* doi:10.1109/UBMYK48245.2019.8965595

Ha, T., Beijnon, B., Kim, S., Lee, S., & Kim, J. H. (2017). Examining user perceptions of smartwatch through dynamic topic modeling. *Telematics and Informatics, 34*(7), 1262-1273. doi:https://doi-org.ezproxy.jyu.fi/10.1016/j.tele.2017.05.011

Ignatow, G., & Mihalcea, R. (2017). Topic models. *Text mining: A guidebook for the social sciences* (pp. 156-162). Thousand Oaks, California: SAGE Publications, Inc. doi:10.4135/9781483399782

Ilgun, K., Kemmerer, R. A., & Porras, P. A. (1995). State transition analysis: A rule-based intrusion detection approach. *IEEE Transactions on Software Engineering, 21*(3), 181-199. doi:10.1109/32.372146

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review . *ACM Computing Surveys (CSUR), 31*(3), 264-323. doi:10.1145/331499.331504

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666. doi:10.1016/j.patrec.2009.09.011

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255. doi:10.1126/science.aaa8415

Karl, A., Wisnowski, J., & Rushing, W. H. (2015). A practical guide to text mining with topic extraction. *WIREs Computational Statistics, 7*(5), 326-340. doi:10.1002/wics.1361

Kemmerer, R. A., & Vigna, G. (2002). Intrusion detection: A brief history and overview. *Computer, 35*(4), 27-30. doi:10.1109/MC.2002.1012428

Ko, C., Ruschitzka, M., & Levitt, K. (1997). Execution monitoring of security-critical programs in distributed systems: A specification-based approach. *Proceedings. 1997 IEEE Symposium on Security and Privacy (Cat. no.97CB36097),* , 175-187. doi:10.1109/SECPRI.1997.601332

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31*(3), 249-268.

Li, J., Qu, Y., Chao, F., Shum, H. P. H., Ho, E. S. L., & Yang, L. (2019). Machine learning algorithms for network intrusion detection. In L. F. Sikos (Ed.), *AI in cybersecurity* (pp. 151-179). Cham: Springer International Publishing. doi:10.1007/978-3-319-98842-9_6 Retrieved from https://doi.org/10.1007/978-3-319-98842-9_6

Lunt, T. F. (1993). A survey of intrusion detection techniques. *Computers & Security, 12*(4), 405-418. doi:10.1016/0167-4048(93)90029-5

M. Esmaili, B. Balachandran, R. Safavi-Naini, & J. Pieprzyk. (1996). Case-based reasoning for intrusion detection. *Proceedings 12th Annual Computer Security Applications Conference,* , 214-223. doi:10.1109/CSAC.1996.569702

McCallum, A. (2002). MALLET: A machine learning for language toolkit [computer software]. http://mallet.cs.umass.edu:

Mills, K. (2017). What are the threats and potentials of big data for qualitative research? *Qualitative Research, 18* doi:10.1177/1468794117743465

Mishra, P., Varadharajan, V., Tupakula, U., & Pilli, E. S. (2019). A detailed investigation and analysis of using machine learning techniques for intrusion detection. *IEEE Communications Surveys & Tutorials, 21*(1), 686-728. doi:10.1109/COMST.2018.2847722

Mukherjee, B., Heberlein, L. T., & Levitt, K. N. (1994). Network intrusion detection. *IEEE Network, 8*(3), 26-41. doi:10.1109/65.283931

Nelimarkka, M. (2019). Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa: Kriittisiä havaintoja . *Politiikka, 61*(1), 6-33. Retrieved from https://journal.fi/politiikka/article/view/79629

Nikolenko, S., Koltsov, S., & Koltsova, O. (2015). Topic modelling for qualitative studies. *Journal of Information Science, 43* doi:10.1177/0165551515617393

Patcha, A., & Park, J. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks, 51*(12), 3448-3470. doi:10.1016/j.comnet.2007.02.001

Phadke, A., Kulkarni, M., Bhawalkar, P., & Bhattad, R. (2019). A review of machine learning methodologies for network intrusion detection. *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC),* , 272-275. doi:10.1109/ICCMC.2019.8819748

Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications, 97*, 205-227. doi:10.1016/j.eswa.2017.12.020

Purhonen, S., & Toikka, A. (2016). "Big datan" haaste ja uudet laskennalliset tekstiaineistojen analyysimenetelmät: Esimerkkitapauksena aihemallianalyysi tasavallan presidenttien uudenvuodenpuheista 1935–2015. *Sosiologia, 53*(1), 6-27. Retrieved from

https://www.researchgate.net/publication/299286511_Big_datan_haaste_ja_uudet_laskennalliset_tekstiaineistojen_analyysimenetelmat_esimerkkitapauksena_aihemallianalyysi_tasavallan_presidenttien_uudenvuodenpuheista_1935-2015_The_challenge_of_big_data_and

Ramage, D., & Rosen, E. (2009). Stanford topic modeling toolbox [computer software]

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* , 45-50. doi:10.13140/2.1.2393.1847

Rohani, V. A., Shayaa, S., & Babanejaddehaki, G. (2016). Topic modeling for social media content: A practical approach. *2016 3rd International Conference on Computer and Information Sciences (ICCOINS),* , 397-402. doi:10.1109/ICCOINS.2016.7783248

S. E. Smaha. (1988). Haystack: An intrusion detection system. *[Proceedings 1988] Fourth Aerospace Computer Security Applications,* , 37-44. doi:10.1109/ACSAC.1988.113412

Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research methods for business students* (8th ed.). Harlow: Pearson. Retrieved from https://www.dawsonera.com/abstract/9781292208794

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms* Cambridge University Press.

Shashank, K., & Balachandra, M. (2018). Review on network intrusion detection techniques using machine learning. *2018 IEEE Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER),* , 104-109. doi:10.1109/DISCOVER.2018.8673974

Sinclair, C., Pierce, L., & Matzner, S. (1999). An application of machine learning to network intrusion detection. *Proceedings 15th Annual Computer Security Applications Conference (ACSAC'99),* , 371-377. doi:10.1109/CSAC.1999.816048

Sisäministeriö. (2017). *Tietoverkkorikollisuuden torjuntaa koskeva selvitys.* ().Sisäministeriö. Retrieved from http://julkaisut.valtioneuvosto.fi/handle/10024/79866

Tabassum, A., Erbad, A., & Guizani, M. (2019). A survey on recent approaches in intrusion detection system in IoTs. *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC),* , 1190-1197. doi:10.1109/IWCMC.2019.8766455

Taylor, S. J., Bogdan, R., & DeVault, M. (2015). *Introduction to qualitative research methods : A guidebook and resource*. Hoboken: John Wiley & Sons, Incorporated. Retrieved from http://ebookcentral.proquest.com/lib/jyvaskyla-ebooks/detail.action?docID=4038514

Toman, M., Tesar, R., & Jezek, K. (2006). Influence of word normalization on text classification. *Proceedings of InSciT, 4*, 354-358.

Tsai, C., Hsu, Y., Lin, C., & Lin, W. (2009). Intrusion detection by machine learning: A review. *Expert Systems with Applications, 36*(10), 11994-12000. doi:10.1016/j.eswa.2009.05.029

U. Lindqvist, & P. A. Porras. (1999). Detecting computer and network misuse through the production-based expert system toolset (P-BEST). *Proceedings of the 1999 IEEE Symposium on Security and Privacy (Cat. no.99CB36344), *, 146-161. doi:10.1109/SECPRI.1999.766911

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management, 50*(1), 104-112. doi:10.1016/j.ipm.2013.08.006

Witten, I. H., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann. Retrieved from http://search.ebscohost.com.ezproxy.jyu.fi/login.aspx?direct=true&db=nlebk&AN=351343&site=ehost-live

Xie, P., & Xing, E. P. (2013). Integrating document clustering and topic modeling. *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, *, 694-703.

Yu, Z., & Tsai, J. J. -. (2011). *Intrusion detection : A machine learning approach*. London : Singapore ; Hackensack, NJ: Imperial College Press ; Distributed by World Scientific Pub. Co. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&AN=373215

Zamani, M. (2013). Machine learning techniques for intrusion detection.

Zhang, H., Kim, G., & Xing, E. (2015). Dynamic topic modeling for monitoring market competition from online text and image data. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, *, 1425-1434. doi:10.1145/2783258.2783293

Zolanvari, M., Teixeira, M. A., Gupta, L., Khan, K. M., & Jain, R. (2019). Machine learning-based network vulnerability analysis of industrial internet of things. *IEEE Internet of Things Journal, 6*(4), 6822-6834. doi:10.1109/JIOT.2019.2912022

LIITE 1 TOPIC EVOLUTIONS OF ALL THE TOPICS

This appendix combines all the topic evolutions.

TABLE 7 Topic evolution of the topic deep learning

1986	2007	2013	2019
learning	learning	learning	learning
data	data	data	data
machine	machine	machine	machine
model	network	network	model
network	model	model	deep
training	deep	deep	network
deep	training	neural	neural
neural	neural	big	training
example	big	training	big
big	supervised	method	method

TABLE 8 Topic evolution of the topic 2 (hard to interpret)

1986	2007	2013	2019
attack	attack	attack	attack
system	system	system	system
security	security	security	security
cyber	cyber	cyber	cyber
control	control	control	threat
cloud	cloud	cloud	control
threat	threat	threat	cloud
network	network	network	network
attacker	attacker	attacker	proposed
based	based	based	based

TABLE 9 Topic evolution of the topic vehicle

1986	2007	2013	2019
detector	detector	detector	vehicle
grid	grid	grid	message
message	message	vehicle	grid
vehicle	vehicle	message	detector
safety	safety	safety	safety
time	time	time	driver
self	self	driver	time
driver	driver	bus	bus
bus	bus	self	vehicular

vehicular	vehicular	vehicular	road

TABLE 10 Topic evolution of the topic intrusion detection system

1986	2007	2013	2019
detection	detection	detection	detection
intrusion	intrusion	intrusion	intrusion
system	system	network	network
network	network	system	system
id	id	id	id
attack	attack	rate	rate
based	based	based	attack
rate	rate	attack	based
false	false	false	proposed
result	result	proposed	accuracy

TABLE 11 Topic evolution of the topic pattern regocnition

1986	2007	2013	2019
recognition	recognition	recognition	recognition
immune	immune	pattern	pattern
pattern	pattern	immune	immune
programming	programming	theory	theory
theory	theory	programming	object
object	object	object	student
evidence	evidence	evidence	programming
multiple	student	student	image
student	multiple	image	evidence
image	image	multiple	multiple

TABLE 12 Topic evolution of the topic internet of things

1986	2007	2013	2019
system	system	system	iot
application	application	data	data
data	data	research	research
research	research	application	application
paper	paper	paper	system
information	information	information	device
security	security	security	security
technology	technology	technology	paper
provide	provide	device	technology
based	device	provide	computing

TABLE 13 Topic evolution of the term 7 (hard to interpret)

1986	2007	2013	2019
mining	mining	mining	user
data	data	data	data
user	user	user	mining
rule	rule	rule	rule
fraud	fraud	study	study
study	study	social	social
information	information	information	web
web	web	web	information
technique	technique	technique	profile
social	social	used	technique

TABLE 14 Topic evolution of the topic 8 (hard to interpret)

1986	2007	2013	2019
method	method	algorithm	feature
classification	algorithm	method	method
algorithm	classification	data	algorithm
data	data	classification	data
result	proposed	feature	proposed
proposed	feature	proposed	classification
feature	result	result	result
classifier	based	based	based
based	classifier	set	performance
set	set	performance	accuracy

TABLE 15 Topic evolution of the topic network attack detection

1986	2007	2013	2019
network	network	network	network
traffic	traffic	traffic	traffic
attack	attack	attack	packet
based	based	based	attack
packet	packet	packet	based
detection	detection	detection	flow
flow	protocol	protocol	detection
protocol	flow	flow	protocol
node	node	node	node

paper	paper	service	service

TABLE 16 Topic evolution of the topic authentication

1986	2007	2013	2019
agent	agent	agent	agent
system	system	system	system
authentication	authentication	authentication	authentication
action	action	action	action
multi	multi	multi	multi
visual	visual	visual	visual
biometric	biometric	biometric	eye
eye	eye	eye	biometric
user	user	user	monitoring
monitoring	monitoring	monitoring	electricity

TABLE 17 Topic evolution of the topic wireless technologies

1986	2007	2013	2019
sensor	sensor	sensor	sensor
wireless	wireless	wireless	wireless
node	node	node	node
proposed	proposed	proposed	proposed
based	based	based	based
algorithm	algorithm	algorithm	algorithm
wsns	wsns	wsns	mdpi
object	object	licensee	switzerland
licensee	licensee	object	basel
network	network	switzerland	licensee

TABLE 18 Topic evolution of the topic particle swarm optimization

1986	2007	2013	2019
optimization	optimization	optimization	optimization
algorithm	algorithm	algorithm	algorithm
swarm	swarm	swarm	swarm
problem	problem	problem	problem
parameter	parameter	parameter	model
particle	particle	pso	particle
pso	pso	particle	parameter

model	model	model	pso
search	search	search	search
fusion	fusion	fusion	fusion

TABLE 19 Topic evolution of the topic anomaly detection

1986	2007	2013	2019
data	data	data	data
anomaly	anomaly	anomaly	anomaly
time	time	time	detection
detection	detection	detection	time
approach	approach	behavior	behavior
event	event	event	approach
behavior	behavior	approach	event
real	real	real	real
stream	stream	stream	stream
log	log	pattern	pattern

TABLE 20 Topic evolution of the topic game theory

1986	2007	2013	2019
trust	trust	trust	trust
game	game	game	game
strategy	strategy	strategy	strategy
trusted	trusted	trusted	ransomware
ransomware	ransomware	ransomware	trusted
risk	risk	risk	equilibrium
equilibrium	equilibrium	equilibrium	member
member	member	member	risk
motif	motif	study	phase
study	study	motif	study

TABLE 21 Topic evolution of the support vector machine

1986	2007	2013	2019
svm	svm	svm	svm
vector	vector	vector	vector
machine	machine	machine	machine
support	support	support	support
kernel	kernel	kernel	signal
signal	signal	signal	kernel
function	function	function	classification
classification	classification	classification	based

based	based	based	accuracy
feature	feature	accuracy	feature

TABLE 22 Topic evolution of the topic 16 (hard to interpret)

1986	2007	2013	2019
model	model	model	model
area	area	study	study
study	study	area	area
result	result	result	result
test	test	test	spatial
spatial	spatial	spatial	test
high	high	high	high
map	map	map	index
document	document	document	map
index	index	index	author

TABLE 23 Topic evolution of the topic 17 (hard to interpret)

1986	2007	2013	2019
model	model	model	model
network	network	network	prediction
prediction	prediction	prediction	network
neural	neural	neural	power
power	power	power	system
system	system	system	neural
time	time	time	energy
parameter	parameter	parameter	time
artificial	artificial	artificial	parameter
forecasting	forecasting	energy	artificial

TABLE 24 Topic evolution of the topic image classification

1986	2007	2013	2019
image	image	image	image
region	region	disease	disease
disease	disease	region	patient
medical	medical	medical	classification
using	using	using	medical
classification	classification	classification	using
patient	patient	patient	diagnosis
brain	brain	classifier	classifier
classifier	classifier	brain	region

diagnosis	diagnosis	diagnosis	cancer

TABLE 25 Topic evolution of the topic 19 (hard to interpret)

1986	2007	2013	2019
domain	domain	domain	domain
gene	gene	gene	source
source	source	source	gene
expression	expression	expression	expression
spectral	spectral	ontology	spectral
study	ontology	spectral	ontology
ontology	study	study	study
e	e	e	e
disease	disease	disease	recurrent
experiment	experiment	recurrent	gru

TABLE 26 Topic evolution of the term mobile malware detection

1986	2007	2013	2019
malware	malware	malware	malware
call	call	call	analysis
method	method	malicious	malicious
malicious	malicious	method	method
program	program	detection	detection
detection	detection	program	feature
analysis	analysis	analysis	call
source	source	source	code
code	code	code	android
use	use	use	technique

TABLE 27 Topic evolution of the topic fuzzy logic

1986	2007	2013	2019
fuzzy	fuzzy	fuzzy	fuzzy
rule	rule	rule	rule
model	model	model	human
knowledge	knowledge	human	model
human	human	knowledge	knowledge
memory	system	system	system
system	memory	decision	decision
logic	logic	logic	logic
decision	decision	cognitive	cognitive
cognitive	cognitive	memory	complex