A natural human language framework for digital forensic readiness in the public cloud

Stacey O. Baror*, Hein S. Venter* and Richard Adeyemi

DigiForS Research Group, Department Computer Science, University of Pretoria, Pretoria, South Africa

*Correspondence to: Stacey O. Baror, email: stacey.baror@cs.up.ac.za;
Hein S. Venter, email: hventer@cs.up.ac.z

ABSTRACT

Currently, about half of all global enterprises are adopting and using some form of cloud computing services. In cloud computing, potential digital evidence is distributed across multiple isolated virtual machine instances. Investigating deleted or inactive virtual instances of a cloud is a challenge to digital forensics, and the traditional methods of digital forensics are inadequate to address such digital forensic investigation. Users of the public cloud (whether a potential victim of a cyberattack, a cybercriminal or a digital forensic investigator) inherently communicate using natural human language in the form of sentences and semantics in document messaging such as texts, emails or instant messages. Consequently, natural human language interaction provides a unique identifier for cloud users. This study leverages the natural human language as an identifier to develop a novel digital forensic readiness (DFR) framework for cloud computing to detect cybercrime. The DFR framework comprises the integration of natural language processing techniques in designing a process that mimics a near real-time approach towards cybercrime detection in a cloud environment. Natural language understanding techniques are used to analyse textdata of users in the public cloud and textdata of reported cybercrimes to develop a DFR framework. In the preliminary formation of the DFR framework, the output shows that cybercrime attacks that are in progress in the form of textdata such as online documents, instant messages or emails within an organizational cloud domain can be identified, and potentially investigated swiftly, using the unique signature of users as identifiers. When adopted, the proposed DFR framework can minimize the time lapses in incident identification and reduce the subsequent investigation time of cybercrimes in the public cloud domain.

KEYWORDS: Cybercrime; natural human language; corpus data; cloud computing; semantics; lexicon; digital forensics

1. Introduction

For decades, technology has progressed considerably due to a massive shift towards automated and computerized systems to complete tedious tasks faster, more efficiently and with minimal error. Cloud computing is the delivery of computing services such as storage, servers, databases, networking and software through real-world services like Email over the Internet [1] . The pay-as-you-use concept of cloud computing is one of the reasons why organizations opt for it. The main difference between the two common cloud computing

deployment models (i.e. the public and the private cloud) is the level of user accessibility. For example, the private cloud services are offered in a private network available only to an organization; therefore, user control is enforced, while the public cloud network is available to the public on a pay-as-you-use basis, and the user accessibility control is flexible [2]. Public cloud computing comprises virtual instances also known as virtual machines. The virtual machine sessions are traceable, while the virtual instance is still running. However, after completion of a given instance's session, the information about the instance (actions performed and session ID) is likely to be irrecoverable. Potential digital evidence can, therefore, be irretrievable and destroyed by such a 'release action'. For example, in a scenario where a user has obtained a public cloud account, the user could well commit a cybercrime ranging from bank fraud, cyberbullying or child trafficking. It could be unfeasible to connect the crime to the offender, due to the volatile nature of the cloud instance. This is possible because any user – even an attacker – can pay for and own a cloud instance, as well as orchestrate an attack with the instance, while instantaneously releasing the virtual machine back into the cloud. The cybercrime committed in the scenario described above or in other similar circumstances in public cloud computing is a challenge for digital forensics.

Due to a lack of user control in the public cloud, specific layers of anonymity are introduced. The anonymity of the public cloud is one of the reasons why cybercriminals like to target the public cloud. However, for a cybercrime to occur, there has to be some form of communication using natural human language to ensure user interaction, be it benign or malicious [3]. The problem investigated in this paper, therefore, is the lack of an easy means to speedily identify a cybercrime in progress, and the researcher proposes that the natural human language of the cybercriminal should be used to detect their crimes. To this effect, this paper proposes a digital forensic readiness (DFR) framework and techniques for safeguarding the public cloud. The proposed DFR framework addresses the problem of speedy discovery and identification of cybercrime incidents in the public cloud by leveraging natural language processing (NLP) approaches. One such NLP approach involves natural language understanding (NLU) techniques [4-7]. NLU aspects such as name entity recognition (NER), text summarization and topic modelling techniques are fused into the architecture of sentiment analysis to extract valuable information from cybercrime textdata (i.e. corpus), whether it originates from cybercrime incidents that are in progress or were reported in the past. However, the reactive nature of the classical digital forensic investigation process [8] renders such digital forensic process inefficient when attempting to attribute a cybercrime to an offender in a near-real-time scenario, a process called user attribution [9,10]. There is, therefore, the need to develop a mechanism to connect a cybercrime actor to a cybercrime attack in progress, as well as to collect potential digital forensic evidence of cybercrime in public cloud instances. Thus, this study addresses the challenge caused by the lack of adequate means to swiftly deploy a digital forensics-based solution in the identification of cybercrime in public cloud computing using corpus data. This is so especially when the public cloud instance is no longer in use (i.e. the cloud data resources were released back into the cloud domain by its user). Consequently, using the corpus textdata [11] in the form of natural human communication/messaging language, this study seeks to connect a cybercrime to the cybercriminal in near real-time through the developing of a DFR framework for public cloud computing.

The remainder of Section 2 presents a background overview of digital forensics, cloud computing and NLP. Section 3 summarizes a previous work that shows the taxonomy of cybercrime attack approaches in the public cloud based on three categories. Section 4 introduces a high-level view of the proposed DFR framework. Section 5 evaluates related

literature and clearly highlights the contributions of this study. Section 6 critically evaluates and concludes this paper, while Section 7 points to future work.

In the next section, a concise overview is given of digital forensics, cloud computing and NLP.

2. Background

The background topics below include the definition and discussion of digital forensics, NLP and cloud computing.

2.1. Digital forensics

Digital forensics is a branch of forensic science that is concerned with the preservation, identification, extraction and documentation of digital evidence [12-14]. To ensure conformity to legal standards, the ISO/IEC 27043 standard summarized the legal requirements and the digital forensic investigation's readiness process to be employed in any digital forensic investigation [15-18]. Furthermore, the National Institute of Standards and Technology (NIST) [2] and McKemmish [19] suggest that carrying out digital forensic investigations or preparing an environment to be ready for a potential digital evidence collection or investigation in a cloud computing domain still poses some challenges. However, the DFR framework proposed in the current study attempts to address these challenges in the cloud computing domain while adhering to the preservation of digital evidence integrity, maintaining a chain of custody, obtaining appropriate authorizations and ensuring forensic soundness [20-22]. This is achieved by the integration of NLP techniques into public cloud forensics. A brief overview of NLP is presented next.

2.2. Natural language processing

NLP is a subfield of computer science, information engineering and artificial intelligence, and it is concerned with the language interactions between computers and human (natural) languages [7,23]. To interpret sentences, the machine parses the sentences grammatically and associates the words with things in the real world in the context of the sentence [7]. For a machine to understand what has been typed from the keyboard, interpretation is required. An NLP pipeline is a chain of independent modules of words, where each word is treated as an input to the subsequent module [5-7]. NLP textdata deals with the underlying latent metadata that addresses the frequency counts of words, the length of the sentence, the presence or absence of certain words and the underlying semantics, pragmatics and syntax. All of these are needed for the machine to understand the meaning of a given sentence in a corpus data feature extraction [11]. The NLP pipeline breaks down tasks into subtasks to be solved independently, where the output of one module feeds into the input of the next module. Each task or subtask of NLP can be accomplished by using machine learning (ML) algorithms. An NLP pipeline includes modules such as morphological analysis, semantics, sentiment analysis and similarity in text. Humans can give the machine a text, voice message or document, while the NLP techniques use ML algorithms to derive meaning from human language in sentences. The sentences are split into a smaller piece of characters or words by using speech recognition, natural language generation (NLG) and NLU. Thus, NLP uses the NLU and NLG to formulate and identify the conceptual communication goal in the linguistic structure of a user by using learned features and rules as the building blocks (items) used in language learning [7,24,25]. NLU techniques, such as NER, sentiment analysis, text summarization, aspect

mining and topic modelling, use logical inference to extract valuable information from text, such as store cybercrime report data.

In simple terms, natural language (NL) is the process where a human gives input to a computing system in the form of a sentence to be processed, based on human-understandable languages such as in English or Zulu [3,24,26]. Applying NLP techniques to extract meaningful information about a cybercrime incident that is in progress in a public cloud is the focus of this study. The concepts of cloud computing as they relate to cybercrime are presented next.

2.3. Cloud computing concepts

Definitions of cloud computing were drawn up by the NIST, and the cloud computing definitions of Jansen et al. [2] were adopted in this study. The NIST defines cloud computing as a model for enabling convenient, on-demand computing service delivery and network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or interaction [1,27,28]. Cybercriminals use cloud computing as an environment to conduct illegal activities, for example, distributing malware, conducting scams, identify theft and other criminal activities [29,30]. According to Pichan [30], it is difficult to prevent cybercrime attacks in the cloud since the amounts of data provided by the cloud customers are desirable and irresistible to cybercriminals. Digital evidence collection is the heart of any digital forensic investigation; however, collecting potential digital evidence in a cloud computing environment is time consuming, precisely due to the vast amount of data. In addition, the standard digital forensic tools are sometimes obsolete in a cloud domain.

According to Gartner [31], half of the current global companies may be fully cloud automated by 2021. The essential characteristics of cloud computing are on-demand self-service, broad network access, resource pooling and rapid elasticity, which is measured using three service delivery models, i.e. Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). SaaS is a software service delivery distribution where a third-party provider, also known as cloud service provider (CSP), hosts and delivers software and application services while making them available to users over the Internet. The user subscribes to the software or services and then accesses the services via the web or application programming interfaces (APIs) [28,30,32]. Examples of the SaaS delivery model include SalesForce.com, Google Mail or Google Docs [33,34]. The PaaS offers a software development platform to users. Users are responsible for protecting the applications they build and run on their platform [35]. Such platforms include tools for creating, testing and implementing application software, database management systems, middleware and the programming environment that is given by the cloud provider [33]. The development platform offered by PaaS hosts both completed and in-progress cloud applications. An example of PaaS is Google AppEngine [34]. The IaaS delivery model focuses on the delivery of IT infrastructures such as network resources, storage and computing power to subscribing users. However, the clients are allowed to control operating systems, virtual storage and deployed applications, but they do not have full access to networking components. An example of an IaaS delivery process is Amazon Web Services, which is currently one of the largest IaaS providers. According to Jansen et al. [2], there are three broad classes of public cloud computing models:

(i) The free-of-cost-to-the-users model. The free-of-cost class of the public cloud is supported through advertisements personalized and delivered to the user using the user's information,

online behaviour and websites visited. Services in this class include search engines, document services and online office or desktop application services.

(ii) The fee-payable-with-no-advertisements model. The low cost and the service level agreement (SLA) terms are non-negotiable by the consumers.

(iii) The final class of the public cloud model service is the fee-payable model. The fee-payable services are negotiated and tailored to the consumer's needs. The services provided by public cloud service providers are sometimes vulnerable to common cybercrime attacks in the cloud, which are discussed next.

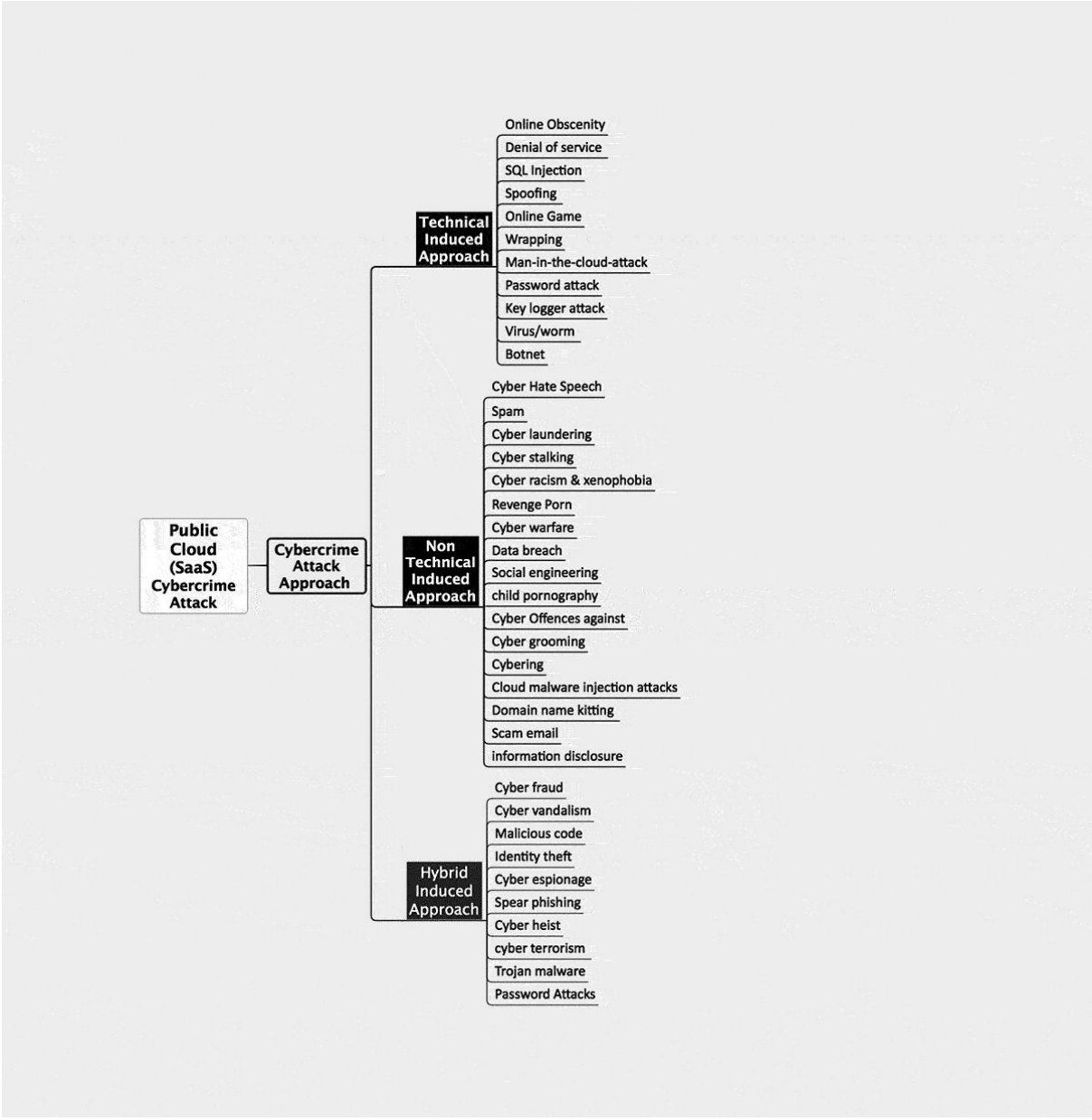

Figure 1. A taxonomy of cybercrime attacks. [3]

3. Previous work on common cybercrime attacks in the cloud

In the last decade, the defence strategy for cybercrime in the cloud has been focused on technically induced cybercrime [3]. In a recent work by Baror et al. [3], an in-depth description of the overall concept of cybercrime induced by natural human language proposes a complete extension towards the realization of a DFR framework to detect cybercrime that is in progress in the public cloud. As shown in Figure 1, Baror et al. [3] classified various cybercrimes based on text and non-textdata used to achieve the cybercrime. The paper identified three categories of cybercrime, namely technical, non-technical and hybrid cybercrime.

The technically induced cybercrime focuses on cybercrime that is perpetrated without prior messaging, email or other text document communication between the attacker and a victim of cybercrime. Such cybercrime attacks include Botnet, SQL injection and man-in-the-middle attacks (see Figure 1). More often than not, when technical cyberattacks occur, the attacks are investigated without taking into account the attacker's use of NL or text interactions. The non-technical cybercrime is cybercrime carried out with the aid of some form of textual communication between the attacker and potential victim. A hybrid cybercrime attack involves the use of tactics that require and combine both technical and non-technical skills of a cybercriminal. The hybrid-induced cybercrime attack approach is a process employed when cyberattack exploits of cybercrime used both technical and non-technical skills, i.e., the attacker's technical skill and text communication were combined to carry out an attack successfully. When both technical and non-technical knowledge is used by a cybercriminal to achieve an attack, it is termed a hybrid-induced cybercrime attack. For comprehensive details of the previous work, see Baror et al. [3]. The next section introduces the proposed DFR framework.

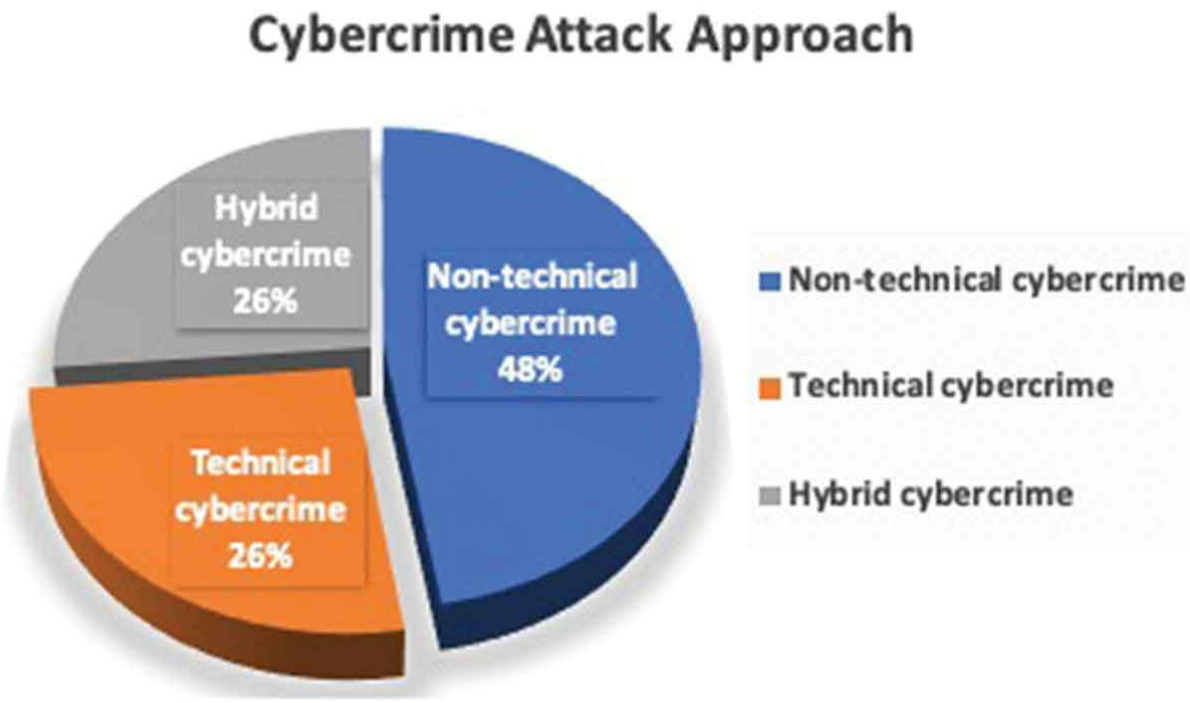

Figure 2. A summary of the taxonomy of cybercrime in the public cloud. [3]

4. A high-level view of the proposed digital forensic readiness framework

Figure 3 depicts a high-level view of the proposed DFR framework. A description of the components of the high-level view of the DFR framework is presented next.

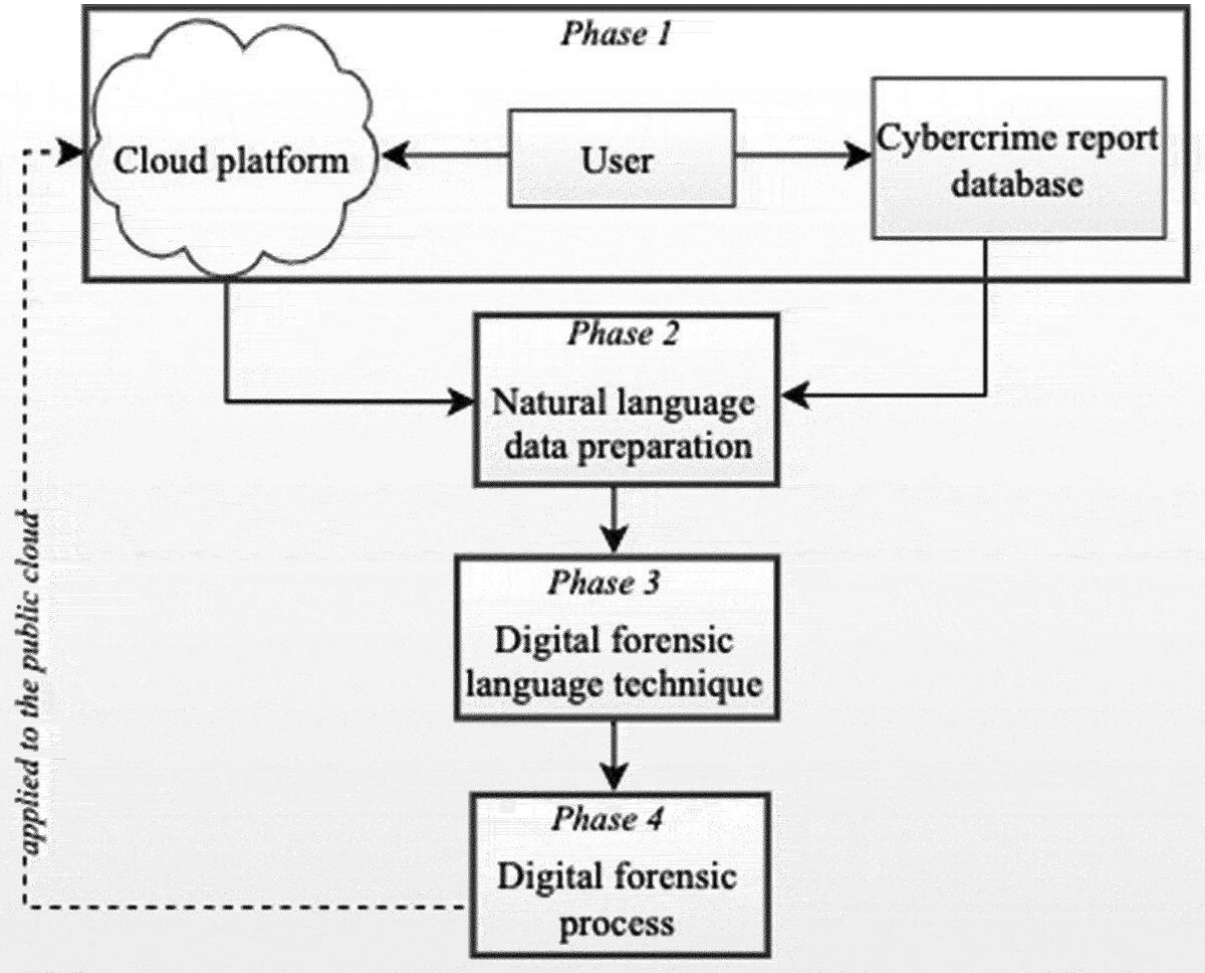

Figure 3. High-level view of the components of the digital forensic readiness framework.

Phase 1: Cloud cybercrime data source: As depicted in Figure 3, this phase consists of the cloud platform component, the user component and the cybercrime report database. Although the simulation of the proposed DFR framework will be carried out in the public cloud, the DFR framework can be adopted to any cloud computing platform.

Component 1: Users: The users, as shown in Figure 3, represent the humans who interact with the various services offered in the cloud domain, such as the SaaS resources which are typically provided by the public cloud as a free-of-cost service. The user could either encounter an attack or be the attacker at some point of their interaction with the system. Details of this component are discussed in Section 4.1.2.

Component 2: Cybercrime report database: This component stores data generated from current and past cybercrime reports by victims of cybercrime attacks. The textdata gathered from the previously reported crimes by victims of cybercrime attacks is passed through the NLP data application component where the analysis of the cybercrime report data is carried out to identify a potential cybercriminal's language pattern. Details of this component are

discussed in Subsection 4.1.3. The data to be analysed must be generated at some point in the DFR framework lifecycle, and the cybercrime report database stores the generated data.

Phase 2: Natural language process (NLP) data application: The NLP component performs the grammatical data analysis of the sentences or phrases of reported cybercrime to extract meaningful textdata for storage and further comparison to the in-progress cloud data. The NLP data application components constitute the composition of the cyberattacks language detection and generation. As depicted in Figure 3, the output of the NLP components is used to develop the cybercrime language detection tool and the digital forensic language library (DFLL). Further details on the NLP components are discussed in Section 4.2.

Phase 3: Digital forensic language (DFL) technique: This component creates the DFL tool using the data derived from the user and the NLP components. The behaviour of the tool will be further examined against the public cloud platform. Details of this component are further discussed in Section 4.3.

Phase 4: Digital forensic process: The digital forensic process components, as depicted in Figure 3, are used by the DFR framework as an incident alert trigger. This trigger action is used to commence the digital forensic process, that is, to trigger the formation of DFR and investigation by using the user's language patterns generated from the various components depicted in Figure 3. Details of this component are further discussed in Section 4.4.

As shown in Figure 3, the components of the DFR framework involve four phases. The specific functions and interactions of the various components of each phase are discussed in the sections that follow.

4.1. Phase 1: Cloud-based cybercrime data source

Figure 4 shows that the cloud-based cybercrime data source consists of a cloud platform, users and a cybercrime report database that generate and store data reported by cybercrime victims. This first phase is the point of interaction between the victims, cyberattackers and digital forensic investigators (DFIs). The semantics and styles of the language of the cyberattackers are compared to textdata from the stored previous cybercrime report and the cybercrime in-progress textdata communication of a cloud instance; these are fused together for the compare-and-detect phase. The data identify cybercrime, and the directional flow shows the data feed to the NL data preparation phase. The various components of the cloud cybercrime data source as depicted in Figure 4 are discussed next.

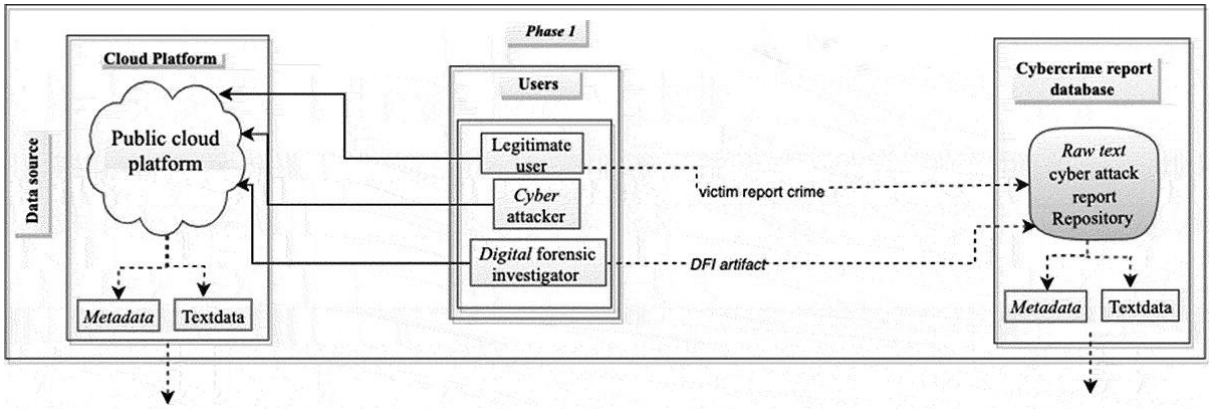

Figure 4. Phase 1 – Cloud cybercrime data sources.

4.1.1 Public cloud platform

The DFR framework hypothesizes that users of the cloud (whether legitimate or cyberattackers) use language such as English as its natural human language. Typically, access to public cloud data requires a complex administrative protocol which may be cumbersome for an individual to carry out. However, using simulation and experiment, this process could be carried out on a public cloud domain that implements the proposed DFR framework. Therefore, the development of the DFR framework will be simulated in the public cloud because the solution architecture and a real-world scenario of the DFR framework are tested in a private cloud. As depicted in Figure 4, the public cloud platform is one of the components of Phase 1. The legitimate user and the cybercrime attackers establish a handshake over the public cloud domain via resource sharing, documents and messaging communications, and the DFI adds digital forensic investigation's artefacts to the cybercrime report database.

4.1.1.1. Metadata

Digital evidence from the cloud is generated from (i) intrusion detection systems; (ii) application and software logs; (iii) cloud service API calls; (iv) system calls from virtual machines [36]; and (v) other available potential digital evidence sources in the event of a cybercrime investigation. This information often includes textdata and metadata. The metadata gives a detailed description of the information about the data (textdata) to identify its aspects and to employ such information in the text analysis stage of the trigger system.

4.1.1.2. Textdata

Textdata is data generated from the grammatical analysis of a sentence. In the context of this paper, the textdata is generated from the reports of cybercrime victims and used as input into the NLP data preparation component of the DFR framework. It is then processed and stored in the cybercrime semantic database builder before being used to compare and determine the semantic validity of textdata from the cloud to identify language patterns. As depicted in Figure 4, when enough textdata and knowledge, words, phrases or sentences are given in an instant message, email or other document message, a cybercrime in progress can be identified. The cybercrime so described has been categorized by a recent research finding [3] as a non-technical cybercrime attack. However, textdata could be read and flagged by the NLP

trigger component based on the user's use of natural human language, without infringing on the privacy of the legitimate users.

4.1.2. User

The user component, as depicted in Figure 4, is the human or person who interacts with the public cloud infrastructures. Many users are adopting cloud computing as the default option for services, applications and software for their individual and business needs. These users are classified as end-users and constitute one of the points of vulnerability in the public cloud environment; therefore, they are most susceptible to cyberattacks. The victims are relied upon to report the attacks in any form, such as by means of reports to the law authorities, news articles, documentary films or blog posts when possible. Other categories of users are the DFI and the potential cybercriminal (see Figure 4). A legitimate user may encounter a cyberattack at some point during the course of their interaction with a cloud instance. The implementation of a framework like the proposed DFR framework could prove effective in the event of a cybercrime attack targeted at human users.

4.1.2.1. Legitimate user

As depicted in Figure 4, the legitimate user in the context of this study is a human person who interacts with or consumes any cloud services. Figure 4 shows a bi-directional flow of users accessing resources of the public cloud. The end-user is an everyday user of the public cloud services that focus on the software as a service (SaaS) platform, where a legitimate user could encounter an attack. Malicious actors (cybercriminals) exploit vulnerabilities of the public cloud to attack public cloud service consumers (eventually the cybercrime victims).

4.1.2.2. Cyberattacker (criminal)

A cyberattacker is an intruder – either a human or a user agent – that accesses the public cloud services with a malicious intent to compromise the legitimate user. All attacks, whether by automated program or human, are classified in this category of user. The cyberattacker attempting to access a legitimate user's account eventually triggers a component of the DFR framework; this further initiates a stylometric trigger when a texting or grooming communication (i.e. non-technical) cybercrime attack approach is detected. The familiar textdata pattern detection forms part of the data-gathering process of the DFR framework. The data generated and collected at this point is used to identify, investigate and prosecute cybercriminals.

4.1.2.3. Digital forensic investigator

In the context of the current study, the role of the DFI as one of the three identified users of the public cloud is shown in Figure 4. In the ISO 27043 process [16-18] for a digital investigation to occur, a readiness process is introduced to include scenario definition, identification of potential digital evidence sources and planning a pre-incident gathering. Although the DFI is an end-user of the cloud, they are tasked with the investigation of crimes post-mortem. The digital forensic investigations employ a chain of custody and evidence validation and also ensure the preservation of digital evidence integrity, adhering to legal requirements, while obtaining proper authorization [13,21,37,38] . Therefore, the NL trigger component of the DFR framework provides an incident scene detection for a DFI. The role of the DFI for the DFR framework also evolves continually throughout the life cycle of the DFR framework.

4.1.3. Raw text cyberattack report repository

As depicted in Figure 4, the raw text cyberattack report repository is a function of the cybercrime report database component. The raw text consists of metadata and textdata that are classified using NLP techniques – beginning with tokenization of the text to the segmentation of words, and punctuation using rules specific to language. The goal of the cybercrime report data repository is to provide textdata to be extracted for semantic meaning using NL techniques. For example, in a typical crime-reporting process, the victim of a cybercrime describes the incident by providing a textual communication and a behavioural description of the attacker [39,40] . Such a description includes the language used by an attacker, the context of the attack, known attributes of the attacker, as well as the perceived vulnerability points explored. These attributes are then analysed and further used to develop a cybercrime report database. The reported cybercrime database prompts the development of an anonymous process for a cybercrime victim to report an attack [41] . The anonymous cybercrime attack-reporting process for the DFR framework serves three purposes:

1. It provides an anonymous cybercrime-reporting sphere.
2. It serves as a data collection, storage and training process for determining the frequency of certain grammar, phrases, sentences and semantics in the use of languages by cybercriminals in a cyberattack.
3. It acts as an input for the application of the ML and NLP techniques in the detection of cybercrime attacks that are in progress in the public cloud.

The cybercrime report data storage provides cybercrime storage for victims to report cybercrime anonymously. It focuses on the collection of raw textdata of cyberattack reports and the detailed accounts of cybercrime incidents reported by victims. The textdata is interpreted to create a filtration point and for data comparison with the live data of the public cloud users. That is, the live cloud dataset is compared to the processed cyberattack report dataset to streamline semantic similarities. The cybercrime report database of the DFR framework is to interface with other cybercrime-reporting databases using microservices and API calls. Cybercrime report databases such as the South African Police Service (SAPS) cybercrime data [42–44] , the United Nations Office of Drugs and Crime (UNODC) under the process of the Commission on Crime Prevention and Criminal Justice (CCPCJ) [45,46] and that asprovided by International Resources on Cybercrime to enhance the volume of the dataset for accuracy. Phase 1 is focused on the potential cybercrime data gathering, while Phase 2 of the DFR framework employs the gathered data using NLP techniques to prepare the data. The NL data preparation process is discussed next.

4.2. Phase 2 – Natural language data preparation

As depicted in Figure 5, the NL data preparation and potential trigger components are used to develop a cybercrime language detection tool and DFLL. They, in turn, create a DF language incident alert to trigger the digital forensic process. Data gathered from the cybercrime victim's communication interaction with a cyber attacker are collected and analysed to become the data comparison with the live cloud data. The NL data preparation phase is used to interpret the contextual user's data such as characters, words, phrases and sentences in a natural human-understandable language [7,47] .

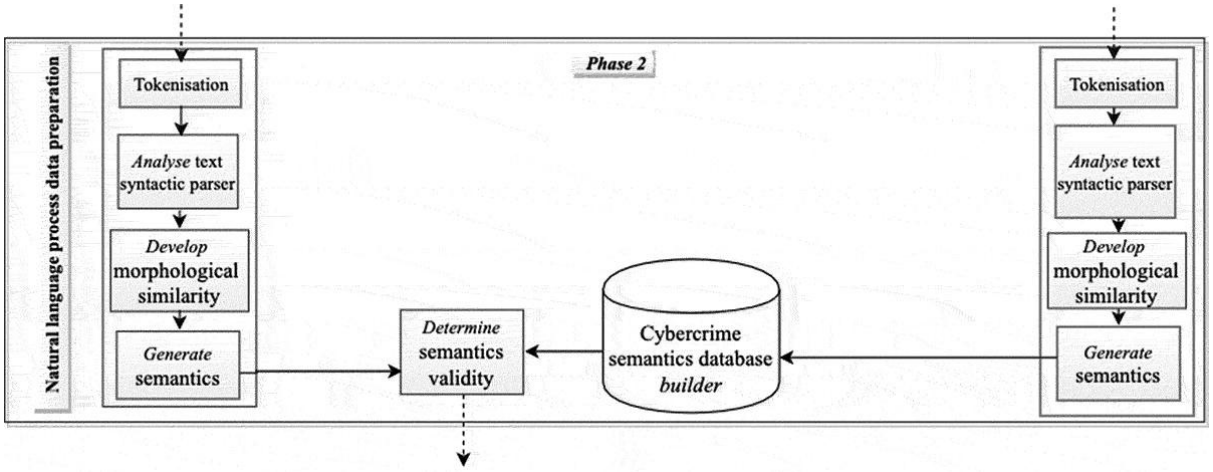

Figure 5. Phase 2 – natural language data preparation process.

As shown in Figure 5, the NL data preparation trigger activities involve the extraction and analysis of human written language in text communication that is transformed into patterns and frequency to automate the identification of patterns in cybercrime communication in the public cloud. These identified patterns are used to develop the following components of Phase 2 of the DFR framework: (i) tokenization; (ii) analysis of text syntactic parse; (iii) development of morphology similarity; (iii) generation of semantics; (iv) determination of semantic validity; and (v) development of a cybercrime semantic database builder (see Figure 5). The patterns of a user's natural human language are addressed using a combination of (i) rule-based; (ii) text summarization; (iii) named entity recognition; (iv) aspect mining; (v) offline and online ML-based; and (vi) hybrid-based text classification model [48,49] . A description of the components of this phase presented in the subsections below includes tokenization, text parsing, as well as morphological and semantic context (and content) generation.

4.2.1. Tokenization

In the context of this study, tokenization is the splitting of text into meaningful segments, called tokens. Tokenization processes are applied to the textdata generated by the NLP trigger from the users of the cloud services who are usually the target of the cyberattacks. The tokenization uses the NL of the user's semantics as an identifier and maps the textdata of the users (such as emails or instant messages) to an attack format. The process also desensitizes textdata extracted from the reported cybercrime storage based on the reports of the various victims of cybercrime attacks.

4.2.2. Develop syntactic parser for text analysis

Syntactic analysis is the process of analysing a string of symbols in natural human language, computer languages or data structures conforming to the rules of a formal grammar [7,50] . The grammatical meaning of a sentence is dependent on the words' structural organization. NL parsing is a process of analysing the complete words and syntactic structure of a sentence. In the DFR framework, a syntactic parser is employed to identify metadata. The textdata is structured following English grammar rules, where the subject or object of a verb in a sentence is used to identify the content to formulate potential cybercrime language. The analysis of the text syntactic parser is used to identify sentiment, context and meaning [50] , to

determine the state of the texts/words/sentences used and to establish how the used words could form a vital rule set for the generation of the digital forensic cybercrime language. Therefore, given pool of textdata, words and phrases of cybercrime, the DFR framework computes the probability that texts, words, phrases or sentences used in a previous cyberattack would be used again in a new cyberattack. Morphology similarities that could identify such texts, words or sentences are discussed next.

4.2.3. Develop morphological similarity

Morphology focuses on the structure of words, prefixes and suffixes. Morphology also deals with how the words are formed and how they relate to other words. In the context of this study and as shown in Figure 5, the goal of the morphological similarity component is to analyse the similarity and to determine the potential of a cyberattacker's use of words that could match to words/phrase in the database. This is based on the use of words in documents, shared content, instant messages and other social interaction textdata used in cybercrime grooming to lure cybercrime victims. The words/phrases of the morphological analysis are used as input into the generated semantics component.

4.2.4. Generate semantics

Semantics is concerned with meanings, word reference (denotation) and associated concepts (connotations). The two main areas of semantics are logical and lexical semantics [7,24,49] . In this section of the cybercrime textdata analysis, the focus is on lexical semantics. Lexical semantics is concerned with the analysis of the meaning of words gathered from the developed morphological similarity components and the relationship between the words. The words formed using the developed morphological similarity are then assembled to form a sensible and identifiable semantics of cybercrime-related words. As depicted in Figure 5 (Phase 2), the semantics process is of two instances – first, using data from cloud platform and second using cybercrime report data from the reported cybercrime database components (see Figure 4 (Phase 1)). Semantics from both the cloud platform and the cybercrime report database components is compared to ascertain commonalities before they are stored in the cybercrime semantic database builder. The cybercrime semantic database builder is discussed next.

4.2.5. Cybercrime semantics database builder

The cybercrime semantic database builder uses the input from the 'generate semantics' component identified in Figure 5 Phase 2 of the DFR framework from the various textdata generated. The cybercrime semantic database builder then employs a lexical semantic analyser to determine the validity of the 'generate semantics' component (see Figure 5). The cybercrime semantics information is analysed to identify consistency and similarities to the semantics data from the reported cybercrime database and that of the cybercrime reports repository, which are capable of parsing the semantics information mined from various reported cybercrimes.

4.2.6. Determine semantic validity

In the context of the DFR framework, ensuring semantic validity is the technique used to further validate the correctness of the semantic textdata generated. Approaches such as the unification of NL inference, common sense reasoning and useful word combinations are

considered to assert the validity of the textdata. The 'determine semantic validity' component is employed to analyse and determine potential false positives in the created and stored cybercrime semantic language data, so as to minimize or eliminate false positives in the DFR framework.

4.3. Phase 3 – DFL technique

As depicted in Figure 6, the DFR framework develops an efficient pro-active cybercrime identification process. Phase 3 (DFL technique) receives its textdata inputs from Phase 2 (i.e. the NL data preparation). Valid semantic data from the cloud cybercrime data sources (Phase 1) imply that the observed semantic characteristics of the instance match known cybercrime semantics in the cybercrime semantic database builder (see Figure 5), which then creates a language trigger process. The DFL technique phase uses the lexical properties (i.e. both logical and lexical) extracted from the semantic composition of the cloud instance. Upon validation of the semantic similarity between the textdata of a new cloud instance and the corresponding cybercrime report textdata in the database, both contents are fused into the cybercrime semantic database builder. The two textdata generation processes (cloud platform lexicon generator and cybercrime lexicon generator) are evoked. Semantics from the input data that originates from the cloud instance is channelled to the first process, while semantics from the cybercrime database is channelled to the second process as shown in Figure 6.

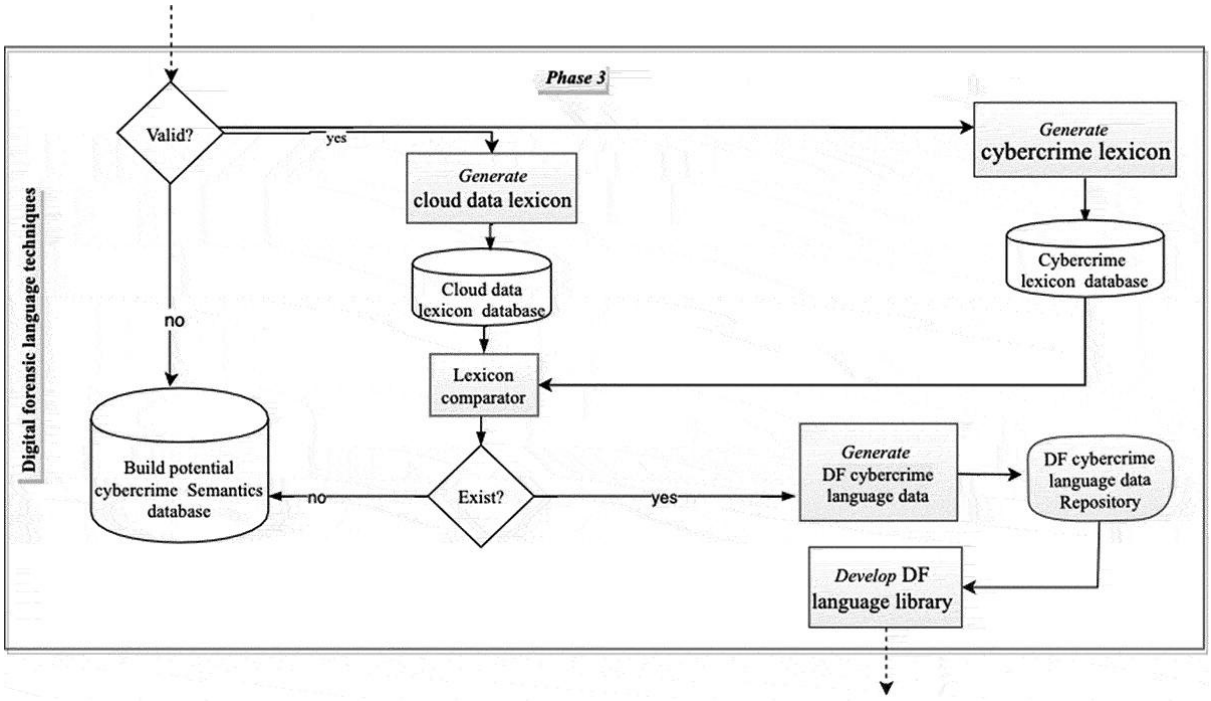

Figure 6. Phase 3 – Digital forensic language technique.

4.3.1. Generate cloud data lexicon

As depicted in Figure 6, the cloud data lexicon extracts lexicon data from the cloud platform (see Figure 4 for Phase 1). The textdata from the cloud platform is used to develop the cloud semantics, which is then used to develop the cloud lexicon of textdata delivered from the cloud platform. To develop the DFL technique phase of the DFR framework, the cloud-based lexicon is used in conjunction with the cybercrime lexicon to achieve a cloud lexicon

comparator (for comparing the cloud and cybercrime lexicon) to extract potential cybercrime language triggers. Classically, a lexicon-based approach towards pattern observation and extraction works on the assumption that the collective polarity of a phrase is the sum polarity of member-words in the phrase. This approach typically depends on a pre-defined corpus of words with respective pre-defined polarity. A cloud lexicon generation process, therefore, involves the development of a novel approach towards lexical identification and extraction. This novel approach comprises the development of a metaheuristic algorithm that can be used to parse the lexicon. The generation of a cybercrime lexicon is discussed next.

4.3.2. Generate a cybercrime data lexicon

The 'generate cybercrime data lexicon' is the second process evoked by the validation process. The semantics from the cybercrime database is parsed as input into this process (see Figure 6). In similitude to the cloud lexicon generator, this process utilizes a metaheuristic algorithm to generate a cybercrime lexicon, which is then used to develop the cybercrime lexicon database. The cybercrime lexicon generation component of the DFR framework is developed as a tool for managing lexicons and generating fraudulent-text dictionaries for an automated lexicon comparison process using cloud lexicon comparator. The cloud lexicon comparator is presented next.

4.4. Lexicon comparator

The outcome of the cloud and cybercrime lexicon generator is fed into the cloud lexicon comparison process (see Figure 6). This comparison process collates the lexical components of the cloud instance with that of the cybercrime database to ascertain the lexical similarity of the cloud instance to that of cybercrime in-progress textdata. The lexical comparison of the DFR framework is also based on the assumption that the observed polarity of a given text is grounded on the overall polarity of the words which produced it. The lexical similarity will, therefore, accommodate lexical complexities associated with natural human language. One potential approach to address the complexities would be to use fine-grained micro-phrases delineated by cues such as punctuation, adverbs and conjunctions. Furthermore, it could include the development of a heuristic technique for the detection of irony, language intensifiers, as well as downtowners. The non-valid semantic comparison serves as an input to consistently grow the potential cybercrime semantic database (see Figure 6, Phase 3 of the DFR framework). The non-valid semantic comparison is applicable in a case where the semantic composition of a given cloud instance does not match any known cybercrime semantics. The outcome of the non-valid semantics is not discarded; instead, it constitutes an input as additional textdata to develop a potential cybercrime semantics database. The latter implies that the stored data could contain probable cybercrime semantics that was not identified at the time of first lexicon comparison. With enough cybercrime textdata, the accuracy of the lexicon comparison increases and heightens the potential of generating real-world DF language data for cybercrime detection (discussed next).

4.4.1. Generate DF cybercrime language data

The generation of DF cybercrime language data is a process to develop language data peculiar to cybercrime. It requires the integration of NLU components, and techniques such as NER, text summarization and topic modelling techniques are fused into sentiment analysis to extract the potential of textdata to build a cybercrime language. To enhance the accuracy of the 'generate DF cybercrime language data' component of the DFR framework, techniques

such as term classification, term frequency and inverse document frequency (used for weighting of lexicons) could be employed. English-specific grammar rules are used for the creation of lexicons to generate cybercrime language. To yield a better granularity with higher discriminative power, a language-specific ad hoc lexicon as identified by Hussein [50] can be used to build a corpus and a cybercrime language detection engine. The DFLL development process is discussed next.

4.4.2. Develop a DFLL

This section creates the proposed DFLL. The 'develop a DFLL' component of the DFR framework is designed to allow for the learning, storing and generation of users' language styles in relation to the language used in cybercrime attacks. The DFLL gathers data, for example, from previous criminal investigations (DFI artefacts – see Figure 4), individual online behaviour and reported physical crimes or cybercrime (cybercrime report database – see Figure 4). The textdata gathered (as depicted in Figure 4) increases the textdata pool and therefore boosts both the potential identification of cybercrime in progress and the learning of the DFR framework in cybercrime language profiling – which leads to the application of the digital forensic process discussed next.

4.5. Phase 4: Digital forensic process

The DFL trigger activates the digital forensic process (see Figure 6) and triggers the alert that identifies a potential cybercrime incident in progress. The output of the trigger is an input to start analysing the data generated; therefore, it triggers the commencement of a digital forensic investigation.

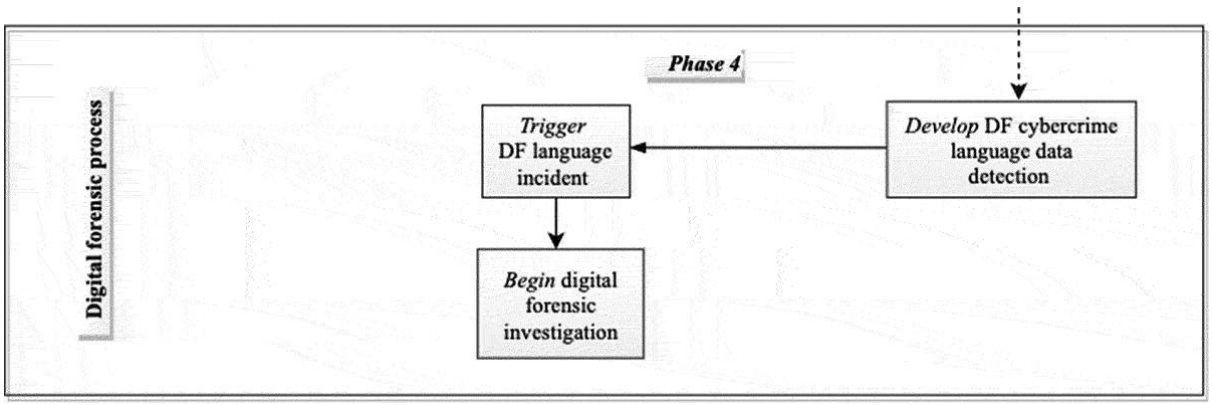

Figure 7. Phase 4 – digital forensic process.

4.5.1. Develop DF cybercrime language data detection

The 'develop DF cybercrime language data detection' process involves the development of a cybercrime language data detector that uses textdata of previous cybercrime reports and of a cloud instance parser. The textdata is compared and categorized in the appropriate class; that is, the lexical characteristics of the cloud platform textdata are subjected to feature parser engineering, pattern recognition and extraction. Feature engineering typically entails data transformation, feature definition, feature identification, feature extraction and feature generation [49,51,52] . Feature engineering is essential to address challenges (such as the cause of dimensionality) that are often associated with NL data analytics techniques. In essence, the

techniques employed to develop the DF cybercrime data detection should satisfy the following conditions:

1. Proactivity: This entails the capability of the developed technique to support the identification, extraction and processing of lexical data in a forensically sound manner and using an automated process.
2. Reliable accuracy: To satisfy the typical forensic criteria, a high degree of accuracy is essential for the reliability of the technique.
3. Robustness to noise is an essential feature that the DF cybercrime language data detection must ensure. Given the high potential of noise in a NL-based approach towards cybercrime, the developed technique must handle noise efficiently as part of the cybercrime language creation and data extraction.

As depicted in Figure 7, the successful implementation of the DF cybercrime language data detection creates an input for the 'trigger DF language incident' component of the DFR framework. This is discussed next.

4.5.2. Trigger DF language incident

The trigger for the DF language incident creates the patterns that identify textdata communication of cybercriminals in a cloud platform, which is also an input to the digital forensic investigation component as depicted in Figure 7 (see also Figure 4). The result of the 'trigger DF language incident' component consists of inputs from the (i) NL data preparation component, which triggers the process (see Figure 5); from the (ii) cybercrime report database (Figure 4); and from the (iii) cybercrime semantic database builder (Figure 5) – all of which contribute to the generation of textdata that is compared at lexicon comparator component (Phase 3, Figure 6). Comparing the cloud lexicon and the cybercrime lexicon eventually generates the DF cybercrime language data, which in turns develops a DFLL to carry out the 'detections' of cybercrime language. The 'trigger DF language incident' process (see Figure 7) relies on the data language detection of previous cyberattacks, victim reports of cybercrime that uses semantics and lexical compositions common to cybercrime attacks employing common sense inference [53-56] . The output of the 'trigger DF language incident' initiates the digital forensic investigation, which is discussed next.

4.5.3. Begin digital forensic investigation

According to the ISO/IEC 27043 [15,16,21,22] , digital investigation is the use of scientifically derived and proven methods for the identification, collection, transportation, storage, analysis, interpretation or presentation of digital evidence derived from digital sources. In the process, authorization must be obtained for all activities to properly document activities, interact with the physical investigation, preserve digital evidence and maintain the chain of custody. All this is required for the purpose of facilitating or furthering the reconstruction of events found to be incidents requiring a digital investigation, whether of a criminal nature or not. Next, the digital forensic investigation process starts with the identification of a potential cybercrime, using the triggers that arise from the NL data preparation (Phase 3) as depicted in Figure 5. An incident is triggered based on language identification formed from the combination of corpus data from report cybercrime data of victim report (see detailed cybercrime report process in Baror et al. [41]). The DFR framework as presented in this paper uses the digital forensic incident response processes as prescribed by the ISO/IEC 27043 [15–18] to tackle the incident.

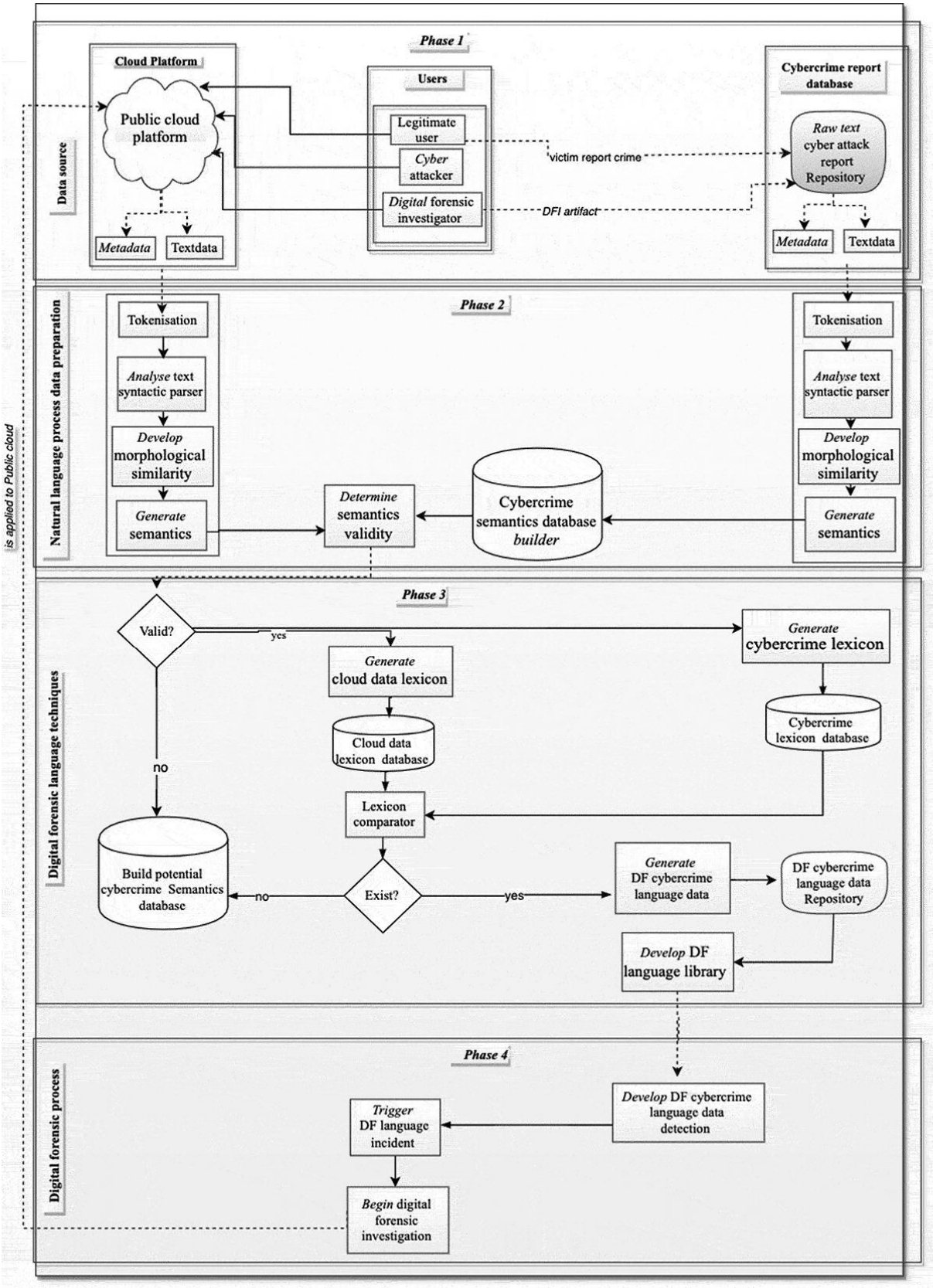

Figure 8. Comprehensive view of the digital forensic readiness framework.

4.6. Comprehensive view of the digital forensic readiness framework

Figure 8 is a detailed illustration of the interaction between the phases and various components of the DFR framework. It provides a detailed view of the natural human language mechanisms that enhance the detection of cybercrime in the public cloud, with emphasis on (corpus data) text-based communications [11] , such as instant messages, document communication, as well as textdata reports of cyberattacks provided by victims. The data generated are used to develop a DFL to trigger an alert whenever a prior cybercrime language pattern is detected (see Figure 8). The next section evaluates literature.

5. Related literature

The purpose of this study is to propose a DFR framework based on NLP techniques that can identify and extract text-based information that constitutes a cybercrime. Other literature related to this issue is discussed next.

Research by Rasmi et al. [57] presents an algorithm that describes the similarities between the different types of cybercrime that could be analysed to identify similarities. It uses multiple attack intention algorithms to find the similarities between cybercrime and then calculates the probability of new cybercrime intentions. These similarities are subsequently recorded and used to predict similar intentions in future cybercrime. Analysing cybercrime intentions gives an in-depth view of what to look for during a cybercrime investigation and therefore predicts the intentions of future cybercrime. The research conducted by Rasmi et al. [57] is likened to this study because the purpose of this paper is to identify the similarities between users' communication styles and use of language. It employs NL process techniques, which, when analysed, identify potential cybercriminals in the public cloud by using word and sentence patterns. It also designs a high-level DFR process that can be triggered based on the user's language pattern during a cybercrime attack in the public cloud.

Heartfield et al.'s [58] motivation for developing the taxonomy of cyberthreats to smart homes is to establish a systematic means for classifying attack vectors and the impact of the cyberthreats within the smart home domain. The authors classified cyberthreats taxonomy based on Internet of Things (IoT). The study investigated attack vectors that resulted in technology convergence in the household and examined the physical, domestic and emotional impact of these attacks on users of the smart homes. Heartfield et al. [58] also considered the potential impact of attacks on smart home systems and on the people utilizing smart home amenities by basing their classification on 25 smart home attacks. The research of Heartfield et al.'s [58] is one of the motivations for the proposed DFR framework adopted by this study to introduce a process that addresses cybercrime in the public cloud.

Alex et al. [59] describe what is needed for a digital forensics investigation to be successful and how cloud providers should implement their systems to help DFIs find the data that they require to find who committed the cybercrime. The cloud provider should also be able to help the investigators access the data in a lawful way to keep its integrity intact and ensure that whatever evidence the investigators find can be used in a court of law. It should be possible to obtain the evidence in a way that guarantees that it has not been modified. In their study, Alex et al. [59] focus on how to ensure a DFR domain, and they give insight into what is required for a domain to be ready for a digital forensic investigation. Their study describes what procedures need to be followed for digital evidence to be valid in a court of law and suggests how cloud providers can ensure that the domain is ready for a potential digital

forensic investigation, therefore guaranteeing a speedy investigation when required. In line with the finding of Alex et al. [59], the current paper focuses on making provision for DFIs using the DFR framework for the public cloud domain as a starting point. When fully implemented, the proposed DFR framework will be able to obtain potential digital evidence in a forensically sound manner, thereby enhancing the validity of the potential digital evidence in any court of law.

Manoj et al. [60] proposed a model based on a trusted third party (TTP) that includes the cloud forensic investigators as a solution to enhance trust in the cloud environment, therefore providing a means to collect digital evidence that could connect an attacker to an attack. The model proposed by Manoj et al. [52] should address the challenge of cybercrime attack investigations in the cloud. In line with the proposal of Manoj et al., the current paper seeks to propose a DFR framework that embeds the mechanisms of gathering potential digital evidence and preparing the cloud environment for potential cybercrime attacks (i.e. ensuring its forensic readiness). Furthermore, the DFR mechanism explored in this paper focuses on using the cyberattacker's natural human language to trace patterns of their previous text communications that resulted in victims reporting such attack. The mechanism also employs the techniques of NLP to isolate the text communication to act as a trigger that identifies a cyberattack while in progress.

Vassil et al. [37] argue that, in the context of evidence acquisition and analysis techniques, the cloud forensics problem requires a new approach of the forensic toolset. This toolset focuses on interfacing directly with the CSP, while addressing the needs of private and public cloud APIs. The forensic toolset addresses issues such as acquisition, analysis and screening of cloud data. It also provides capabilities that combine the insight of the client and the service provider to achieve DFR in the cloud. The DFR framework proposed in this study focuses on designing toolsets to address the challenges of cloud computing by using natural human language interactions and communications to detect and investigate cybercrime in the public cloud domain. The current study employs the victims, the trusted third party (DFIs) and the public cloud environment to activate the process that generates the data required for the DFR frame. For future studies, the CSP interface, as suggested by Vassil et al. [37], will be integrated to enhance the data collection process.

6. Evaluation

The proposed DFR framework enables the timely identification of a potential cybercrime in progress in a near real-time manner. The near real-time capability of the proposed DFR framework is achieved by analysing the textdata gathered from the cybercrime report data, from data generated as a result of users' interactions in the public cloud delivery model (e.g. SaaS), artefact data from the DFI and other sources. This research was motivated by the adoption of cloud computing by small, medium and large-scale organizations as core its infrastructure. Atrend that has exponentially increased in the last decade. For example, according to Gartner and other researchers,[31,61,62] the global spending on cloud services has grown from just over 46 billion USD in 2008 to 260 billion USD in 2018 (a 465% increase in 10 years).

The problem that this study attempted to address was the lack of adequate means to address DFR that swiftly identifies a cybercrime in progress in the public cloud, especially when the virtual cloud instance is no longer in use. The proposed solution to this problem is to implement a natural human language DFR framework as a trigger for the public cloud to

detect cybercrime attacks in a near real-time manner. The DFR framework harnesses the text communication of the cyberattacker and the report of victims of the attacks to develop a semantic and lexicon builder for cybercrime language. The proposed DFR technique is useful, novel and unique because, for cybercrime to occur, there must be communication (direct or indirect) that uses language. In the case of direct communication, the natural human communication language could be French, English, Zulu, etc. The core component of the interaction between cybercriminals is communication. Such communication is used by public cloud users to interact with other humans or non-human agents, whether they are benign or malicious. However, the reactive nature of the classical digital forensic investigation (DFI) process renders it inadequate when attempting to connect a cybercrime to the offender in real time. Furthermore, this paper seeks to address the problem of the speed of cybercrime attribution in a public cloud platform by developing a DFR framework for the public cloud computing platform. A typical public cloud platform generates and uses metadata, text, sentences and semantics via documents, emails or instant messages. These textdata plus data from cybercrimes report of cyberattack victims are used to develop a DFLL process. Another main challenge of addressing cybercrime, irrespective of the cloud computing platform, is the inadequacy in reporting of cyberattacks by the victims. Victims of a cyberattack are sometimes unable to anonymously report cyberattacks which inherently deter them from reporting such attacks. The victims of a cyberattack are often portrayed as a criminal. So much so that the victims or organization will preferably not be associated to a cyberattack. Therefore, the victimization of cybercrime victims has further resulted in a lack of user-based objective data to tackle cybercrime. This notion of low cybercrime reporting creates a new challenge that has the following results: (i) insufficient data to build a predictive textdata cybercrime detection system; (ii) the cybercriminal evading legal prosecution; (iii) an insufficient knowledge base for both research and business purposes; and (iv) the non-existence of human language interaction and communication as a DFR process.

A recent publication [3] made a classification of various cybercrime attacks using NL to identify unique attack vectors. The study asserted that in every successful cybercrime attack, there is a 61% chance that the cyberattacker communicated in a language such as English, French, German, etc. Despite the above findings, cyberwarfare focused to a considerable extent over the last decades on addressing cybercrime attacks that were technically induced, such as by means of ransomware, SQL injection, denial-of-service and other technically sophisticated cyberattacks related to data breaches. Very little attention has been accorded to the text-based, non-technical communication that often typifies cybercrime attacks. The current paper identified the need for a digital forensic process that focuses only on text communication, the natural human language of the users (victims, criminals and investigators alike) and previous cybercrime attack data, as a prompt to commence a DFR process in an organization.

However, implementing any DFR process, especially for the public cloud, faces a series of challenges from economic factors, post-mortem (after-the-fact) cybercrime incident identification/detection, privacy, secrecy/trust and the stigma of cybercrime reportage. Other challenges include international legal jurisdiction and tenancy issues (i.e. where the potential evidence is located across the world), privacy issues, SLAs and, above all, customers' service delivery to the end-users, coupled with the cost to the public CSPs. The proposed DFR framework for the public cloud based on human NL unobtrusively checks the activities of end-users and allows the public cloud services to place a flag on events that are suspected to be crime related. This approach has the potential to address the cybercrime in-progress identification challenges. Furthermore, the DFR framework provides economic incentives to

stakeholders (i.e. end-users; the public CSPs; the potential cybercrime victims; and the DFIs) such that the cost of cybercrime identification and investigation is reduced by a certain margin since a mitigation process that identifies potential cybercrime is in place. This is based on the assertion that a large percentage of the costs of cybercrime is not driven by the number of cybercrime incidents, but by the bad experience of the affected organization and how the organization chooses to protect its business interests to maintain its investors/market position [63,64]. The DFR framework proposed by this research also addressed the post-mortem challenges of cybercrime (i.e. after-the-fact cybercrime incident identification/detection), because one of the component of the DFR framework generate cybercrime data that could be used for the swift identifications of what happened and how it happened [65-67], in any digital forensic investigations. Volatile information can be collected while the instances are still active. The analysis of the corpus of previous cybercrime reports presents an opportunity to identify cybercrime attacks in near real time. This approach leverages the knowledge-based technique to interactively identify, extract and present a similar pattern from the corpus of previous cybercrime reports.

The proposed DFR framework for the public cloud is essential for global cyber safety proposed by various national and international organizations. This assertion is captured in the 2010 United Nations (UN) report on cybersecurity, which announced that the cybercrime identification timeline is a significant global challenge [45,46,68-70]. The UN General Assembly recognized the real and significant risks posed by attempts to identify and detect cybercrime long after the cybercrime incident had occurred [70]. The inability to tackle cybercrime in-progress is a global challenge that has, in a way, advanced cybercrime. Therefore, a speedy cybercrime incident identification process is one of the contributions of this paper.

Privacy constraint is a challenge that could impede profitability of the public cloud, whereby some CSPs tend to rebuff DFR. This challenge transcends to the proposed DFR framework. Privacy is one of the reasons some users opt for cloud computing service to preserve and retain their anonymity. The DFR framework could act as a trusted third party in the preservation of cyberlaws and maintenance of DFR in the cloud computing environments. Trusted third party is in many ways are presented as the intermediary to act in the interest of two or more parties [71,72], in this case the CSP and the users. In legal terms, trust in records is dependent on the knowledge of (i) the creator or custodian of the records; (ii) the reputation of the creator or custodian of the record; (iii) past performance and competence of the creator or custodian of the record; and (iv) the assurance of confidence in future production. Therefore, for the DFR framework proposed by this study, uphold the legal requirements and trust agents put in place to act on behalf of a legitimate public cloud user (i.e. a potential cybercrime victim), the DFI and the public CSP.

7. Conclusion

In conclusion, this paper describes a DFR framework for gathering digital information that can be used for DFR in the public cloud. The study focuses on the cyberattacker's use of everyday natural human languages as an identifier to detect a cyberattack in progress. It is further asserted that developing a digital forensic framework – based on users' use of natural human language – provides a trigger for swift identification and investigation of cybercrimes in the public cloud. The research suggests a four-phased process that flows from the cloud-based and cybercrime-reported data source (Phase 1) to the NLP data preparation (Phase 2), to DFL techniques (Phase 3) and finally to the digital forensic process (Phase 4). The three components in Phase 4 interact to achieve the entire process of using natural human language

to identify cybercrime incidents in public cloud computing. The application of digital forensics in the cloud computing platform is not without its challenges, and the proposed DFR framework is merely the first phase of this study.

Future work will explore the process of generating the 'unstructured natural human language data' in real time from emails, chats, web pages or social media of a public cloud platform service models. Applying an unsupervised machine and online learning to the stored reported cybercrime textdata and cloud platform data will be used to develop a prototype of the DFR framework. The prototype will first develop a digital forensic cybercrime language library (DFLL) to act as a beacon for the near real-time cybercrime detection. Second, the DFLL optimizes the identification and investigation of cybercrime using digital forensic techniques.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

1. Fox, A. et al. Above the clouds: A berkeley view of cloud computing. Dept. Electr. Eng. Comput. Sci. Univ. California, Berkeley, Rep. UCB/EECS; Pg1-26, 2009.

2. Grance T, et al. Guidelines on security and privacy in public cloud computing? Public Cloud Comput Secur Priv Guidel. 2012:1–95. doi:10.3233/gov-2011-0269.

3. Baror OS, Venter H A taxonomy for cybercrime attack in the public cloud. In: 14th International Conference on Cyber Warfare and Security, ICCWS 2019. 505–515; 2019.

4. Roukos S. Natural language understanding. Springer Handbooks Pearson; 2008. doi:10.1007/ 978-3-540-49127-9_31.

5. King BE, Reinold K. Natural language processing. Find Concept. 2014:67–78. doi:doi:10.1016/ b978-1-84334-318-9.50005-3.

6. Liddy E. D. Natural language processing. In: Encyclopedia of library and information science. Marcel Decker, Inc; Surface.syr.edu; 2001. p. 1–15.

7. Hirschberg J, Manning CD. Advances in natural language processing. Science. 2015;349:.261–266. doi:10.1126/science.aaa8685.

8. Vincze EA. Challenges in digital forensics. Police Pract Res. 2016;17:183–194. doi:10.1080/ 15614263.2015.1128163.

9. Ikuesan AR, Venter HS. Digital behavioral-fingerprint for user attribution in digital forensics: are we there yet? Digit Investig. 2019;30:73–89. doi:10.1016/j.diin.2019.07.003.

10. Mohlala M, Ikuesan AR, Venter HS User attribution based on keystroke dynamics in digital forensic readiness process. In: 2017 IEEE Conf. Appl. Inf. Netw. Secur. AINS 2017, Miri, Sarawak Malaysia on 13–14 November 2017- Janua. 1–6; 2018.

11. O'Day DR, Calix RA Text message corpus: applying natural language processing to mobile device forensics. In: Electron. Proc. 2013 IEEE Int. Conf. Multimed. Expo Work. ICMEW 2013; San Jose, CA, USA2013. doi:10.1109/ICMEW.2013.6618380

12. Carrier B. Open source digital forensics tools: the legal argument. Techreport, Stake; 2002.13. Carrier BD. Digital forensics works. IEEE Secur Priv. 2009;7:26–29. doi:10.1109/MSP.2009.35.

14. Arthur KK, Venter HS. An investigation into computer forensic tools. Issa. 2004;1(1):1–11.

15. Valjarevic A, Venter HS. A comprehensive and harmonized digital forensic investigation process model. J For Sci. 2015;60:1467–1483. doi:10.1111/1556-4029.12823.

16. Omeleze S, Venter HS Testing the harmonised digital forensic investigation process model-using an Android mobile phone. In: 2013 Information Security for South Africa – Proceedings of the ISSA 2013 Conference; 2013. doi:10.1109/ISSA.2013.6641063

17.Valjarevic A, Venter HS. Introduction of concurrent processes into the digital forensic investigation process. Aust J For Sci. 2016;48:339–357. doi:10.1080/00450618.2015.1052754.

18. Mumba ER, Venter HS Testing and evaluating the harmonized digital forensic investigation process in post mortem digital investigations. In: Proceedings of the Conference on Digital Forensics, Security and Law. 83–98; South Africa; 2014.

19. McKemmish R. When is digital evidence forensically sound? IFIP international federation for information processing. Boston, MA: Springer; 2008. p. 285.

20. Martini B, Choo KKR. An integrated conceptual digital forensic framework for cloud computing. Digit Investig. 2012;9:71–80. doi:10.1016/j.diin.2012.07.001.

21. Omeleze S, Venter HSHS. Proof of concept of the online neighbourhood watch system. Lect Notes Inst Comp Sci. 2016;171:78–93.

22. Omeleze S, Venter HSHS. Digital forensic application requirements specification process. Aust J For Sci. 2019;51:371–394. doi:10.1080/00450618.2017.1374456.

23. Harris MD. Introduction to natural language processing. USA: Reston Publishing Co; 1985.

24. Strawson PF. Subject and predicate in logic and grammar. Routledge; 2017. doi:10.4324/ 9781315242132.

25. Allen J. Natural language understanding. Addison Wesley, Benjamin/Cumming; 1995, ISBN: 0805303340, 9780805303346.

26. Wilks Y. Natural language processing. Commun ACM. 1996;39:60–62. doi:10.1145/ 234173.234180.

27. Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I, et al. A view of cloud computing. Commun ACM. 2010;53:50–58. doi:10.1145/ 1721654.1721672.

28. Uschie. History & evolution of cloud computing, 2018, associate Vice President in Seasia Infotech, a CMMi Level 5 certified software development organization in its office in Mohali, Punjab. [Online]. 2018. https://www.seasiainfotech.com/blog/ history-and-evolution-cloud-computing

29. Pasquale L, Hanvey S, McGloin M, Nuseibeh B, Pasquale L, Hanvey S. Adaptive evidence collection in the cloud using attack scenarios. Comput Secur. 2016;59:236–254. doi:10.1016/j. cose.2016.03.001.

30.Pichan A, Lazarescu M, Soh ST. Cloud forensics: technical challenges, solutions and compara-tive analysis. Digit Investig. 2015;13:38–57. doi:10.1016/j.diin.2015.03.002.

31. Moore S, van der Meulen R. Gartner survey says cloud computing remains top emerging business risk. Alertify; 2018. https://alertify.eu/gartner-survey-says-cloud-computing-remains-top-emerging-business-risk/

32. Kulkarni G. Cloud computing-software as service. Int J Cloud Comput Serv Sci. 2012;1:11.

33. Arutyunov VV. Cloud computing: its history of development, modern state, and future considerations. Sci Tech Inf Process. 2012;39:173–178. doi:10.3103/S0147688212030082.

34. Zawoad S, Dutta AK, Hasan R. Towards building forensics enabled cloud through secure logging-as-a-service. IEEE Trans Depend Secur Comput. 2016;13:148–162. doi:10.1109/ TDSC.2015.2482484.

35. Abdullah S, Abu Bakar KA, Abbas H, Maennel O, Assar S. Security and privacy challenges in cloud computing. Proc 2018 Cyber Resil Conf. 2019;8:24–31.

36. Liu C, Singhal A, Wijesekera D. Identifying evidence for cloud forensic analysis. In: Research advances in cloud computing; 2017. p. 371–391. doi:10.1007/978-981-10-5026-8_15.

37.Garfinkel S, Farrell P, Roussev V, Dinolt G. Bringing science to digital forensics with standar-dized forensic corpora. DFRWS 2009 Annu Conf. 2009;6:S2—-S11.

38. Prayudi Y, SN A. Digital chain of custody: state of the art. Int J Comput Appl. 2015. doi:10.5120/19971-1856.

39.Finklea KM, Theohary CA. Cybercrime: conceptual issues for congress and U.S. law enforce-ment. In: Cybercrime: conceptualized and codified. Washington, DC: Library of Congress, Congressional Research Service; 2013. p. 1–27.

40. Brown CSD. Investigating and prosecuting cyber crime: forensic dependencies and barriers to justice. Int J Cyber Criminol. 2015;9:55–119.

41. Omeleze Baror S, Ikuesan, Adeyemi R, Venter HS A defined digital forensic criteria for cybercrime reporting. In: Brian P, Hongyi W (eds.),ICCWS20 – Proceedings of the 15th International Conference on Cyber Warfare and Security, USA, 658. ACPIL; 2020.

42. Gould C, Burger J, The NG. SAPS crime statistics: what they tell us–and what they don't. South Afr Crime Q. 2014;42:3–12. doi:10.4314/sacq.v50i1.43. Van Niekerk B. An analysis of cyber-incidents in South Africa. AfrJ Inf Commun. 2017;2017:113–132. doi:10.23962/10539/23573.

44. Morgan S 2017 Cybercrime report. 2017 Cyberrime report. 1–14; 2017.

45. Bilder RB, Zagaris B, Clark RS. The United Nations crime prevention and criminal justice program: formulation of standards and efforts at their implementation. Am J Int Law. 1997;91:408. doi:10.2307/2954230.

46. Clark RS. The United Nations crime prevention and criminal justice program: formulation of standards and efforts at their implementation. Pennsylvania, USA: University of Pennsylvania Press; 1994.

47. Zhou L, Burgoon JK, Nunamaker JF, Twitchell D. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. Gr Decis Negot. 2004;13:81–106. doi:10.1023/B:GRUP.0000011944.62889.6f.

48. Hussein NH, Khalid A. A survey of cloud computing security challenges and solutions. Int J Comput Sci Inf Secur. 2016;14:52–56.

49. Sammons M, Christodoulopoulos C, Kordjamshidi P, et al. Edison: feature extraction for NLP, simplified. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 4085–4092, Portorož, Slovenia; 2016.

50. Hussein DMEDM. A survey on sentiment analysis challenges. J King Saud Univ Eng Sci. 2018;30:330–338.

51. Liu H. Feature engineering for machine learning and data analytics. Feature engineering for machine learning and data analytics. ' O'Reilly Media, Inc.'; 2018. doi:10.1201/9781315181080.

52. Arneklev BJ, Grasmick HG, Bursik RJ. Evaluating the dimensionality and invariance of 'low self- control'. J Quant Criminol. 1999;15:.307–331. doi:10.1023/A:1007528515341.

53. Angeli G, Manning CD NaturalLI: natural logic inference for common sense reasoning. In: EMNLP 2014–2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference; 2014. doi:10.3115/v1/d14-1059

54. Zellers R, Bisk Y, Schwartz R, Choi Y SWAG: a large-scale adversarial dataset for grounded commonsense inference. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 93–104: Association for Computational Linguistics; 2019. doi:10.18653/v1/d18-1009

55. Badecker W, Caramazza A. Morphological composition in the lexical output system. Cogn Neuropsychol. 1991;8:.335–367. doi:10.1080/02643299108253377.
56. Wang Z, Mi H, Ittycheriah A Sentence similarity learning by lexical decomposition and composition. In: COLING 2016-26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers, Osaka, Japan; 2016.

57. Rasmi M, Jantan A, Al-mimi H A new approach for resolving cyber crime in network forensics based on generic process model. In: The 6th International Conference on Information Technology, Amman, Jordan; 2013.

58. Heartfield R, Loukas G, Budimir S, Bezemskij A, Fontaine JRJ, Filippoupolitis A, Roesch E. A taxonomy of cyber-physical threats and impact in the smart home. Comput Secur. 2018;78:398–428. doi:10.1016/j.cose.2018.07.011.

59. Alex ME, Kishore R. Forensics framework for cloud computing. Comput Electr Eng. 2017;60:193–205. doi:10.1016/j.compeleceng.2017.02.006.

60. Manoj SKA, Bhaskari DL. Cloud forensics – a framework for investigating cyber attacks in cloud environment. Procedia Comput Sci. 2016;85:149–154. doi:10.1016/j.procs.2016.05.202.

61. Costello K, Gartner forecasts worldwide public cloud revenue to grow 17.5 percent in 2019. Gartner.Com. 2019:1–5. https://www.gartner.com/en/newsroom/press-releases/2019-04-02-gartner-forecasts-worldwide-public-cloud-revenue-to-g

62. Ramgovind S, Eloff MM, Smith E The management of security in cloud computing. In: Proceedings of the 2010 Information Security for South Africa Conference, ISSA 2010. 1–7; 2010. doi:10.1109/ISSA.2010.5588290

63. Anderson R. Measuring the cost of cybercrime. In: The economics of information security and privacy. Springer; 2013. p. 265–300. doi:10.1007/978-3-642-39498-0_12.

64. Gañán CH, Ciere M, Van Eeten M Beyond the pretty penny: the economic impact of cybercrime. In: ACM International Conference Proceeding Series. 35–45; 2017. doi:10.1145/3171533.3171535

65. Wiles J, Reyes A. The best damn cybercrime and digital forensics book period. The best damn cybercrime and digital forensics book period. Printed in the United States of America 1234567890 and Published by Syngress; 2011. doi:10.1016/B978-1-59749-228-7.X0001-X.

66. Duren BM, Hosmer C Can digital evidence endure the test of time? In: Proceedings of the Digital Forensic Research Conference, DFRWS, Syracuse, New York, USA; 2002.

67. McKenna B. Symantec's Thompson pronounces old style IT security dead. Netw Secur. 2005;2005:1–3. doi:10.1016/S1353-4858(05)00194-7.

68. Satolla D, Judy HL. Towards a dynamic approach to enhancing international cooperation and collaboration in cyber security legal frameworks: reflections on the proceedings of the workshop on cybersecurity legal issues at the 2010 United Nations Internet Governance Forum. William Mitchell Law Rev. 2011;37:1745–1804.

69. Wolter D. The UN takes a big step forward on cybersecurity. Arms Control Today. 2013;43:25–29.

70. Schjølberg S, Ghernaouti-Hélie S. A global protocol on cybersecurity and cybercrime. Oslo: Cybercrimedata; 2011. ISBN 978-82-997274-3-3978-82-997274-3-3; 2e édition2e.

71. Zissis D, Lekkas D. Addressing cloud computing security issues. Futur Gener Comput Syst. 2012;28:583–592. doi:10.1016/j.future.2010.12.006.

72. Schneier, B, Secrets and lies: digital security in a networked world [Books]. IEEE spectrum. Vol. 37. Hoboken, New Jersey: John Wiley & Sons; 2005