Preventing Data Loss by Harnessing Semantic Similarity and Relevance

Hanan Alhindi[1], Issa Traore[2], and Isaac Woungang[3*]
[1]King Saud University, Riyadh, Saudi Arabia
halhindi@ksu.edu.sa
[2]University of Victoria, Victoria, BC, Canada
itraore@ece.uvic.ca
[3]Ryerson University, Toronto, ON, Canada
iwoungan@cs.ryerson.ca

Received: February 15, 2021; Accepted: April 20, 2021; Published: May 31, 2021

Abstract

Malicious insiders are considered among the most dangerous threat actors faced by organizations that maintain security sensitive data. Data loss prevention (DLP) systems are designed primarily to detect and/or prevent any illicit data loss or leakage out of the organization by both authorized and unauthorized users. However, exiting DLP systems face several challenges related to performance and efficiency, especially when skillful malicious insiders transfer critical data after altering it syntactically but not semantically. In this paper, we propose a new approach for matching and detecting similarities between monitored and transferred data by employing the conceptual and relational semantics, including extracting explicit relationships and inferring implicit relationships. In our novel approach, we detect altered sensitive data leakage effectively by combining semantic similarity and semantic relevance metrics, which are based on an ontology. Our experimental results show that our system generates on average relatively high detection rate (DR) and low false positive rate (FPR).

Keywords: Data loss prevention, Threat actors, Malicious insiders, Similarities, Data leakage, Detection rate

1 Introduction

A malicious insider is an authorized member of the organization (e.g., a disgruntled employee) that intentionally and negatively affects the integrity, confidentiality, or availability of the organization's resources including data, system, or network [35]. One of the considerable threats that malicious insiders could cause is data leakage. Sensitive data loss or leakage causes enormous damage to the organization including business functions disruption, financial losses, reputation and brand damage, and regulatory violation fines [10] [47]. According to Campbell, 94% of companies that are victim of data loss don't survive [16]. Data loss prevention (DLP) systems help mitigate insider threats [39] [9]. DLP is a security control that can be used in the organization to monitor, identify, and prevent any illicit data revealed by end-users[39] [9]. Detecting sensitive data leakage is challenging because quite often it is difficult to distinguish these activities from legitimate ones. Also, skilled malicious insiders can leak sensitive data by rewriting it while keeping the same meaning. Furthermore, while detecting structured data format

Journal of Internet Services and Information Security (JISIS), volume: 11, number: 2 (May 2021), pp. 78-99
DOI:*10.22667/JISIS.2021.05.31.078*

*Corresponding author: 350, Victoria street, Toronto, Ontario, M5B 2K3, Canada, Tel: +1-416-979-5000 ext. 6972, Web: https://www.cs.ryerson.ca/~iwoungan/

is straightforward, existing DLP systems struggle with detecting data leakage when unstructured data is involved.

1.1 Research Problem and Contribution

Current DLP systems can monitor and detect sensitive data leakage by matching either file name, file size, or particular sensitive terms contained in a file. Also, these DLP systems focus on detecting and matching structured data since this kind of data can be matched in several easy ways. Moreover, other DLP systems can match unstructured data by fingerprinting the data, which is effective only when the exact copy of the original sensitive data is transferred. However, when the sensitive content is modified, these kinds of systems cannot detect the semantic similarities between the original and modified copies of sensitive data. In this paper, our approach does not focus on structured data detection since this kind of data can be matched and detected easily and directly. However, our approach mainly aims to detect and match unstructured data by using ontology-based semantic metrics including semantic similarity and semantic relevance metrics (introduced later). Furthermore, our approach has the ability to detect an altered copy of sensitive data by using a lexicon of a specific domain of interest. Furthermore, a new dataset in sports domain has been collected, which combines public data under the form of football news articles and private leaked football data.

1.2 Approach Overview

Our new approach detects any attempts at evading detection such as paraphrasing the content while keeping the same meaning of sensitive data. We focus on the document content semantic, which basically consists of both a set of concepts and a set of relationships between these concepts. The extraction of concepts and their relationships uses an ontology of a specific domain of interest. The ontology is the representation of the concepts and the relations among them. Based on the knowledge domain of the organization's data that needs to be protected, we choose corresponding domain ontology for extracting, matching, and detecting semantic similarities and semantic relevance between monitored and sensitive data. The remaining sections of the current paper are organized as follows: Section 2 gives a background on data leakage and loss prevention and introduces some related works on DLP and ontology management and semantic models. Then, Section 3 presents and describes the semantic similarity metrics that have been used in our model. Section 4 presents the datasets and ontology used in our experiments. Next, Section 5 presents the experimental evaluation of the proposed model and discusses its results. Finally, Section 6 makes concluding remarks.

2 Background and related works

2.1 Data Leak Prevention

The increasing online presence of organizations and the growing cyber activities that this entails has led to a steady growth of their data [39]. Having massive amounts of critical data related to the organization, employees, and customers require high internal and external security to exchange data smoothly and safely. As a result of the aforementioned, organizations need to provide easy data communication for their users while protecting and preventing sensitive data from any breach. The breach could be in a form of data loss, which threatens organizations' security, competitiveness, and credibility. Numerous data loss incidents have had negative impacts on the corresponding organizations. For example, the American multinational corporation Morgan Stanley faced an incident of data loss caused by one of its financial advisors called Galen Marsh, who stole critical accounts information of the company's wealth

management clients [23]. As well, before a Walt Disney CEO announced the company's quarterly earnings to the public, he sent an email about that to a reporter by accident in 2000 [29]. Also, a sensitive email about Prince Charles' visit to Poland was sent by mistake to a wrong email address in 2002. As a result of that situation, Prince Charles' plan was disclosed to newspapers before he announced that [29]. Moreover, Marriot International hotel group had experienced a massive data breach since 2004 by unauthorized access to its Starwood database. They reported that around 500 million customers' information were copied and disclosed [46]. All these incidents emphasize that protecting and preventing organizations from data loss is a significant step in terms of security and privacy.

2.2 Data Loss Prevention

DLP is a technique for detecting and preventing unauthorized disclosure, breaches, or exfiltration of the company's confidential information. DLP helps to protect sensitive data by preventing all types of data loss including intentional and unintentional data loss that could affect the confidentiality, integrity, or availability of the organization's sensitive information [21]. Some DLP techniques are used for detection, prevention, or both of the above. DLP techniques often consist of three main elements as shown in Figure 1 [39]:

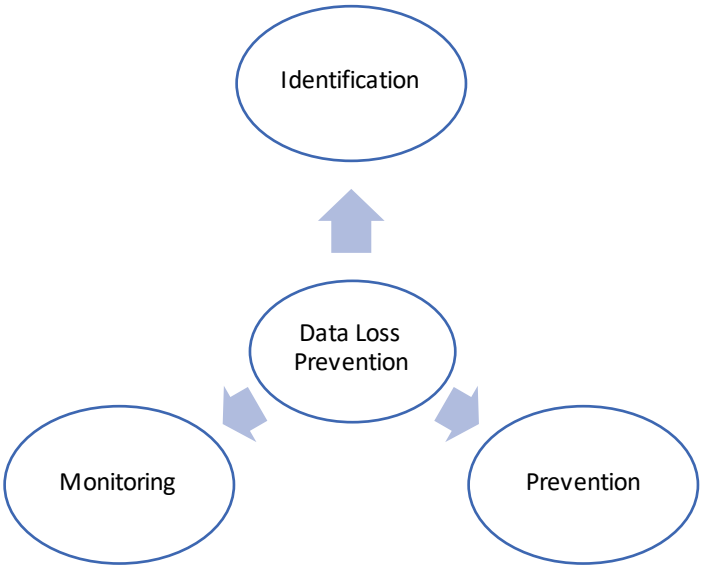

Figure 1: The three main elements of DLP technique [39]

- Identification is a detection method of organizations' critical information based on predefined policies. Whether the DLP technique is just for detection, prevention or both of the above, this method usually consists of an analysis task that is based on either data context, content, or a combination of the above.
- Monitoring is a tracking and flagging method of organizations' critical information that should not leave organizations' networks.
- Prevention is a method of taking actions with flagged critical data that is derived after applying identification and monitoring methods. Different prevention methods are shown in Figure 2 including allowing users to access files or send emails, blocking the access to suspicious files or

emails, labeling users and withdrawing some privileges, reporting the users' activities to the network administration or manager, and warning users about their current suspicious activities. These prevention methods could be applied either individually or in combination when the organization's security policies are violated based on the organization's regulatory requirements.



Figure 2: Several prevention methods of DLP

The organization's confidential information that should be protected and secured can exist in any state based on the data lifecycle as follows [39] [2] [34]:

- Data at rest is the data that is stored on company devices and systems such as databases, servers, file systems, computers, and distributed desktops.
- Data in use is the data that is used and being processed by company users on endpoint devices such as a file being copied to a USB device, a file being accessed in a computer, or an email being written but not yet sent.
- Data in motion is the data that is moved through the company network to the outside world by e-mails, instant messages, web traffic, or other communication methods.

The DLP technique helps to monitor, identify, and protect critical information from a loss or leak outside the corporate network. Data loss may occur through regular storage, usage or transmission data channels. In this situation, it is significant to identify the most popular data loss channels related to data states to provide the desired security level without impeding information flow through those channels. The common data loss channels for data at rest and data in use are through hard copies of documents, removable storage devices, and portable devices. Moreover, data in motion could be leaked while sharing files and information or web services via emails, webmail, instant messages, social media, cloud, and portable devices. Figure 3 shows some examples of common data loss channels.

2.3 Data Loss Causes

There are two main causes of data loss in an organization based on human factors and their location according to an organization:

- External cause: One of the main external causes of data loss is remote attackers. Remote attackers are the outsiders, who do not have access to organizations' data but they can penetrate the system by malware, injected codes, or social engineering attacks to illicitly access sensitive data and cause data loss to an organization [3].

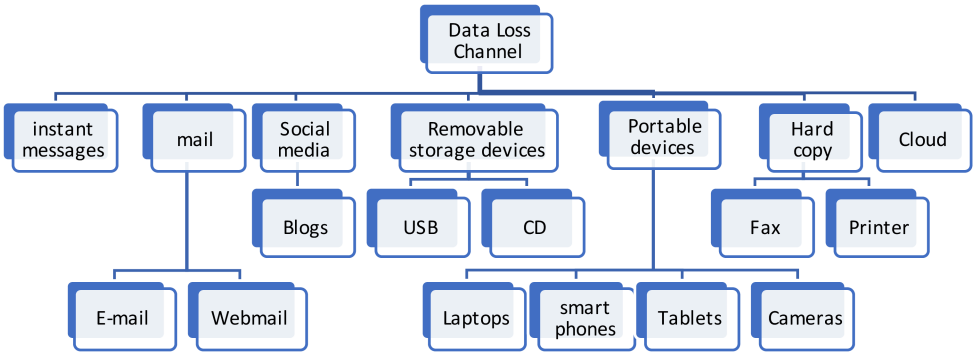

Figure 3: Common data loss channels

- Internal cause: One of the main internal causes of data loss is insider threats. Insider threats are authorized employees, who can intentionally abuse their privileges and maliciously access and transfer sensitive data out of organizations' network boundaries. Those insiders can transfer critical information to outside organizations' networks by either of two ways:
  - *electronically* by sending information via web or e-mails
  - *physically* by sending information via data storage devices such as USBs and hard disks.

Malicious Insider threats could cause data loss in two forms as follows [34] [3]:

- Damaging critical information, which compromises the availability or the integrity of data by either hiding or corrupting the original correct copy of data.
- Leaking critical information, which compromises data confidentiality by unauthorized disclosure.

2.4 Data Loss Categorization

Data loss incidents could happen in any organization either intentionally or unintentionally.

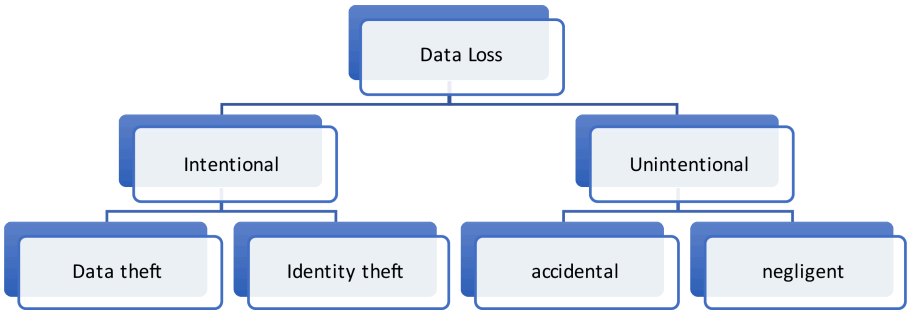

Figure 4: Data loss categorization

To illustrate, insider threats such as disgruntled employees could cause intentional data loss by stealing sensitive organization's data, credit card records, or social security numbers and use them maliciously to destroy the organization's data, system, or finance. On the other hand, unintentional data loss could

happen accidentally by natural disaster and fire, which could harm the system or data. In addition to that, employees could by mistake attach a critical document to an email, send an email to a wrong address, publish or post organization's private information to the public accidentally or negligently. Figure 4 shows the categorization of data loss.

2.5 Data Loss Prevention Approaches

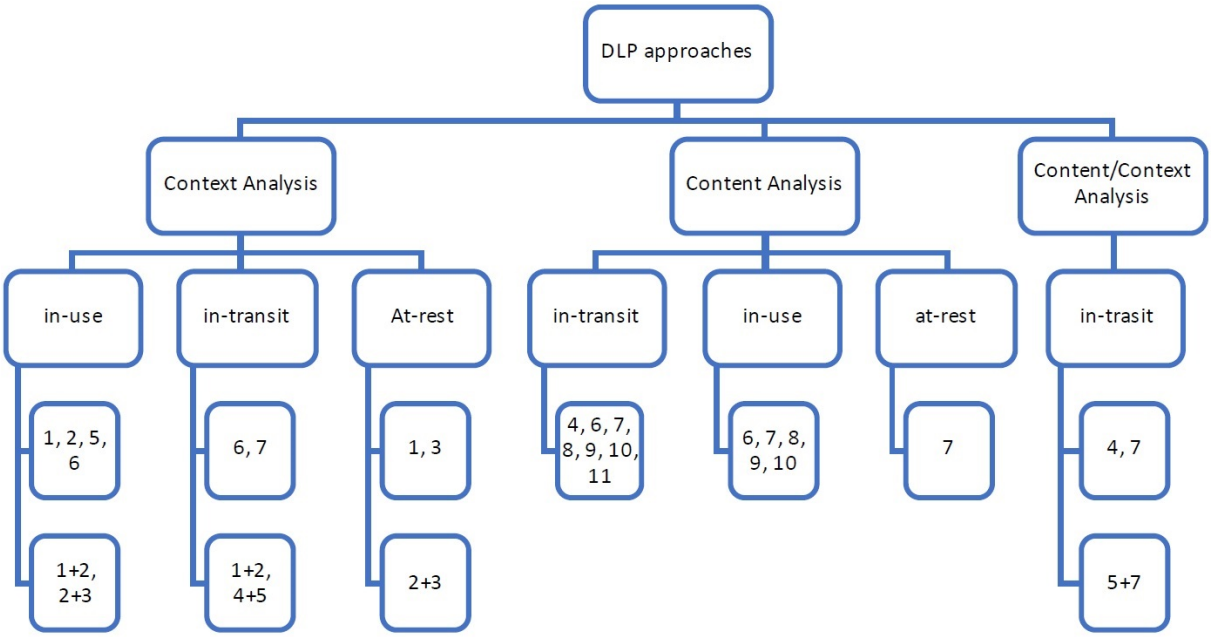

Figure 5: DLP-categorization based on content/context analysis and data states; the numbers refer to the row ids in Table 1

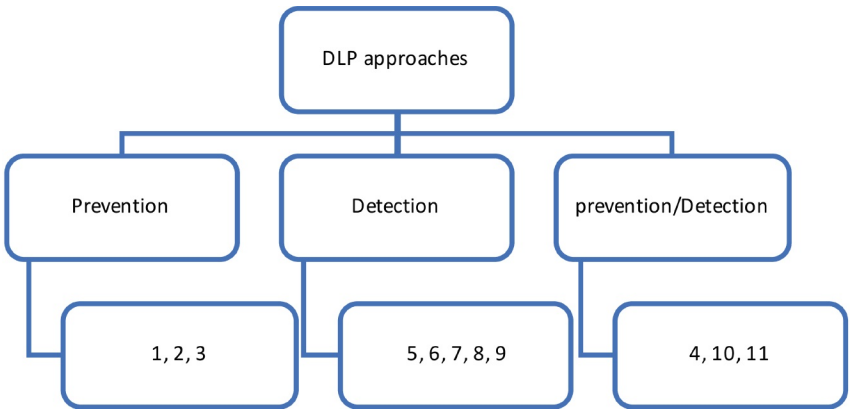

Figure 6: DLP classification based on detection/ prevention methods; the numbers refer to the row ids in Table 1

Table 1: Comparison of DLP techniques.

DLP Id	DLP techniques	Technique goal	Advantages	Disadvantages
1	Policy and access rights [48] [45]	Prevention	- Most common preventive technique for organizations. - Simple to implement and easy to control. - Uses contextual analysis. - One of the robust techniques to prevent sensitive data leak before it happens.	- Only a prevention technique. - Should be combined with a detection technique to detect and prevent sensitive data leak. - When access control and policy is being used, the technique is influenced.
2	Virtualization and Isolation [25] [14]	Prevention	- When a user accesses sensitive information, a trusted virtual environment is created to isolate the user's activities and grant him privileges for certain actions.	- Only a prevention technique. - Should be combined with a detection technique to detect and prevent sensitive data leak. - Is expensive to implement and maintain.
3	Cryptographic approach [11]	Prevention	- Works by generating cipher text from a plain text of sensitive content. - Can prevent users from reading the plain text.	- No guarantee of security for the ciphertext.
4	Quantifying and limiting [19] [49] [13]	Prevention/ Detection	- Able to detect and prevent leaked data. - Administrator utilizes quantifying methods to build and organize critical data. - Interested in leaking channels	- Could interrupt data flow of some channels. - Does not guarantee to fully block leaking channel. - It has limited detection of leaked data via hidden channels.
5	Social and behavior analysis [12]	Detection	- It is considered a proactive prevention technique. - Can detect malicious relationships between people, groups, and organizations. - Can predict and track human behavior.	- Generate a high false positives rate. - Require administrator's monitoring and involvement. - Require collecting users' profiles for comparison and detecting irregular activities.
6	Data identification [42] [28] [27] [43]	Detection	- When this technique uses fingerprinting analysis, it generates a low false positive rate. - Very robust to detect unaltered data. - Can use some powerful hashing technique to detect altered data.	- Prior knowledge of sensitive data is required. - Cannot detect highly altered data. - Cannot understand data semantics.

7	Data mining and text clustering [26] [36] [24]	Detection	- Associated with machine learning techniques - Able to detect sensitive content of unstructured data - It is also a data leak prediction technique. - Able to perform a complicated task.	- May generate a high number of false positives. - Involve a massive amount of computations. - Has limited scalability.
8	Data fingerprinting (exact/partial matching) [42] [43]	Detection	**(Exact data/file matching)** - Uses database for storing exact data or hash values for files - Effective for detecting structured data from the database. - Effective for fingerprinting any type of file. - Produces a low false positive rate. - Uses contextual analysis **(Partial file matching)** - Able to detect sensitive content of unstructured data - Works either for the whole file or part of the file - Generates a low false positive rate. - Uses Content analysis	**(Exact matching)** - The performance is influenced by the large size of the database. - Unable to detect unstructured data. - Unable to detect modified files. **(Partial file matching)** - It is necessary to specify critical documents that need to be protected. - The performance is influenced by the massive volume of critical content that needs to be protected.
9	Regular expression (dictionary-based match) [37]	Detection	- The analysis can be done by comparing the content against specific rules. - It is effective for detecting structured data such as credit card numbers, social security numbers	- It generates a high number of false-positive rates. - It has limited capability to detect sensitive information within unstructured data.
10 11	Statistical analysis (N-gram analysis) [44] Statistical analysis (Term weighting) [17]	Prevention/ Detection	- Uses Content analysis - Able to detect sensitive content in unstructured data - May classify the importance of the content based on machine learning techniques.	- Involve a large amount of data - May generate a high amount of false-positives and false negatives.

DLP approaches are categorized based on analyzing data context, content, or a combination of them. We can consider the content like a letter while the context as an envelope.

Context Analysis The DLP techniques, which are based on context analysis, focus on the contextual attributes that are linked to the monitored data, such as source, destination, time, size and format [10]. Examples of context analysis DLP techniques include social and behavior analysis, data identification, and data mining and text clustering.

Content Analysis The DLP techniques, which are based on content analysis, focus on and inspect the content of monitored data whether it is a document or an email to trigger policy violations in an organization. In this case, those DLP techniques are mainly based on three content analysis types, which are data fingerprinting, regular expression, and statistical analysis [8].

Table 1 presents and compares several DLP techniques by highlighting the underlying goals, and advantages and disadvantages [10] [41] . Figure 5 shows the categorization of data loss prevention approaches based on content and context analysis. Some of these approaches are based on either context, content, or both of them. Also, we classified DLP techniques based on the data that they deal with including data in-use, in-transit, and at-rest. We represented each DLP technique by its id that is shown in Table 1. For example, statistical analysis (Term weighting) approach with id 11 is based on content analysis, which deals with data-in-transit while policy and access rights approach with id 1 is based on context analysis and deals with data-in-use. When policy and access rights approach is combined with virtualization and isolation approach, they analyze data in use and data in transit contextually. Also, data mining and text clustering technique with id 7 could analyze the content and the context of data in transit. Furthermore, Figure 6 classifies DLP techniques into detection, prevention, or detection and prevention techniques representing each technique by its id as shown in Table 1 [10].

2.6 Related works on Data Loss and Leakage Prevention

In 2015, a novel semantic similarity detection approach for data leakage prevention was proposed by Du et al. [22]. This approach uses the latent semantic analysis (LSA) and the support vector machine (SVM) algorithm for sensitive semantic features' extraction to represent concepts. Also, removing stop words and stemming are applied and the relative frequency for each tested document is calculated and compared to a particular threshold to determine semantic similarity and prevent critical data leakage. Du et al. evaluated their model on five document sets and obtained performance rates ranging between 76.1% - 98.6% for TPR and 0.8% - 15.1% for FPR.

Similarly, in our approach, we implemented several steps such as removing stop words, stemming, and calculating the frequency to determine the most representative word in each document in the form of a set of concepts. However, Du et al. introduced the relative frequency by calculating the number of a specific term occurring in a tested document with respect to the number of words in that document. Instead, we introduced the frequency by calculating the number of matched concepts in a tested document with respect to a critical reference document. In addition, our proposed model uses a lexicon of a specific domain of interest, which can provide more concepts' synonyms when extracting terms from tested documents and searching for concepts in an ontology. Moreover, Du et al. used a general dataset, which did not include leaked data and consisted of published research papers from Google scholar, IEEE, and ACM digital library, while we used two real-life datasets that contain critical leaked data.

Shapira et al. proposed an extended method for fingerprinting content to detect any data leakage [42]. The authors focused only on critical contents to generate fingerprints to produce less false alarms. Also, this approach is considered the first research in using k-skip-n-grams for text fingerprinting. In addition, this approach can detect data leakage in modified documents. The evaluation of the approach

was conducted using two datasets consisting of Reuters news articles and a subset of the PAN plagiarism corpus 2010. Under some scenarios, the obtained performance results were acceptable and the accuracy was high. However, there is a need in the approach to improve space efficiency while maintaining a high accuracy level. In addition, the evaluation lacked real-life leaked datasets; this would be required for more representative results.

Costante et al. introduced a hybrid framework for data loss prevention and detection, which includes five main steps: learning, prevention, detection, alert analysis and rule management [20] . Also, the framework aggregates both anomaly-based and signature-based components. Once users' activities are monitored, an anomaly-based component is used to detect transactions with abnormal behaviors. Malicious transactions will be flagged and blocked using a rule-based prevention technique. At the same time, a signature-based component is used to generate an appropriate signature to prevent any upcoming similar transactions. Costante et al. evaluated the framework on two datasets, which are synthetic and real-life datasets obtained from GnuHealth and Oracle, respectively. The authors studied the relation between the number of attacks and time and they found that the framework achieved a quick response in triggering the prevention rules for malicious transactions. However, the accuracy of the approach was not studied; important performance measures, such as detection rate and false positive rate have not been provided.

Kaur et al. proposed a new data leakage prevention algorithm via gateway to protect transferred sensitive emails from any disclosure [30]. The proposed algorithm consists of several steps including creating a sensitive keyword list for each department, checking the email subject, content, and attachment, checking the sender's department, matching extracted keywords from the email with the sensitive keyword list, and preventing the transfer of sensitive email if similarity is determined. The prevention methods that have been introduced are either blocking, encrypting, or quarantining email. The limitation of this method is that it works only with emails but not with other kinds of knowledge containers such as files, blog posts, and repositories. However, our proposed DLP approach supports different kinds of knowledge containers by monitoring and detecting sensitive data leakage from any newly generated content such as new emails being written and new files being created.

Ling et al. proposed a new method for network data leakage prevention by checking TCP data packets [32] . The proposed method stores in a server users' names along with the sensitive information that they are allowed to transfer. Then, the system retrieves the source user information and the sensitive content in any TCP packet being transferred via the network. After that, the system checks if the source user of a transferred packet is eligible to send the sensitive information in that TCP packet. Next, based on user policy the system will either allow or block the transmission of the TCP packet. The limitation of this method is that it allows an authenticated user to transmit TCP packet, which carries a sensitive information; however, the system is not able to detect or prevent sensitive data leakage if an authenticated user discloses sensitive information through that transmission. In other words, the system is able to detect sensitive data leakage if an unauthorized user sends sensitive information via TCP packet, but not when this is done illicitly by an authorized user.

2.7 Related works on Ontology Management and Semantic Models

Li et al. introduced an algorithm to calculate the semantic similarity between two short texts based on semantic nets and corpus statistics [31]. This algorithm uses information content provided by corpus statistics, such as Brown Corpus, to weigh and determine the importance of each word in each sentence. The integration of semantic similarity and order similarity are the basis of calculating the sentence similarity. The calculation of semantic similarity is based on two semantic vectors and information content, while the calculation of order similarity is based on two-word order vectors. Generating two semantic vectors and two word order vectors depend on a lexical database such as WordNet. The algorithm

achieved acceptable Pearson correlation coefficient but it requires other performance measurements to compare it with other methods. Similarly, with our proposed model, we used a lexical database. However, Li et al. used a general lexical database such as WordNet, which may give irrelevant synonyms of the word's context, while we provide a lexicon and an ontology in a specific domain of interest in our model for semantic purposes. In addition, a drawback of Li et al.'s approach is that the generation of semantic vectors and word order vectors for two compared texts is only effective when the text length is short. As a result of that, this approach is not capable to generate the aforementioned vectors and compute the semantic similarity between two long texts.

Liu and Wang proposed a method for assessing text similarity based on an ontology [33] . This method semantically compares between two texts either a sentence or an entire document. The connections that exist between ontology concepts and text content are used to create concept vectors for either sentences or documents. In other words, sentences and documents' concept vectors are generated based on the concepts' weights in a sentence or a document. However, document concept vectors TF_IDF measure to determine words' importance in a specific document. Also, Liu and Wang used semantic similarity indexes, which are derived from the aforementioned connections, for searching and comparison purposes. A comparison is conducted of Liu and Wang method with other methods, such as word overlap, TF_IDF, and linguistic measures. As well, the authors computed four performance measurements in their evaluation, including accuracy, precision, recall, and F-measure, and their method achieved higher values for all four measures compared to other methods.

Chahal et al. developed a semantic similarity model using relation measurements for web documents based on an ontology [18] . The model consists of three main components, namely, ontology processor, document processor, and semantic score calculation components. The ontology processor component consists of concept and relation analyzer, which in turn is used to extract all concepts in a specific web document and relations between them. The document processor consists of a syntactic and a semantic analyzer, which in turn extract words using a specific dictionary in a specific domain. Then, the method calculates the semantic similarity between each pair of concepts in a document by finding the Relation space model (RSM) and the lexical patterns. The RSM has all relations of concepts in a document along with relations' frequencies. The comparison between two documents can be done by comparing the RSM of documents and calculating the semantic similarity score between two documents. The method was evaluated on 50 documents and compared to the vector space model and Euclidean approaches and achieved better performance results compared with other approaches. However, this model needs to be run on a much larger dataset to measure semantic similarities between documents and obtain more meaningful performance results.

3 Ontology-based semantic metrics

To calculate the similarity of the monitored document with respect to the reference document, we use an ontology-based semantic similarity metric and an ontology-based semantic relevance metric defined in the following. Later, in the experimental evaluation section, we will combine these two metrics and assess their performance for our model.

3.1 Ontology-based semantic similarity metric

In our model, we applied a similarity metric to measure the semantic similarities between two concepts based on an ontology. The metric was originally defined in the work of Saad et al. [40] . The semantic similarity between two concepts in an ontology is measured based on the commonalities and the differences between them. In other words, the relations between two concepts and their lowest common

ancestor in an ontology capture the commonalities between these two concepts. In contrast, the locations of two concepts in an ontology capture the differences between them. Given two concepts c_1 and c_2 belonging to documents d_1 and d_2, the semantic similarity between these concepts is defined as follows:

$$sim(c_1,c_2) = 1 - \frac{path(c_1,LCA(c_1,c_2) + path(c_2,LCA(c_1,c_2))}{depth(c_1)+depth(c_2)} \quad (1)$$

Where path(c_1 ,LCA (c_1, c_2)) is the length of the shortest path from concept c_1 to the least common ancestor (LCA) of c_1 and c_2, and depth(c_1) is the depth of concept c_1 in the concept tree. The output of sim (c_1, c_2) is a value between [0,1] where 1 corresponds to exact match while 0 corresponds to no match between the two concepts. We define the document semantic similarity by summing and averaging the semantic similarities between their respective concepts, as defined as follows:

$$sim(d_1,d_2) = \frac{\sum_{c_1 \in d_1, c_2 \in d_2} sim(c_1,c_2)}{n_{d_1} \times n_{d_2}} \quad (2)$$

Where n_{d_1} and n_{d_2} are the total number of concepts in d_1 and d_2, respectively.

3.2 Ontology-based semantic relevance metric

To capture the relevancy between concepts, in our model, we use a semantic relevance metric that depends on the relationships between concepts in an ontology.

Ontology Relations Given an ontology O=(C,R), the concept tree for a document d is defined as the triple $CT(d)$ = { { $Thing$ }, C_d, R_d }. There are two types of relations $r \in R$ between any two concepts in an ontology:

- Explicit relations $r_{explicit}$, which are predefined between classes and need to be extracted from the ontology.
- Implicit relations $r_{implicit}$, which are hidden and need to be inferred by some inference rules.

In our model, we extract the relations between the classes whether they are explicit or implicit relations. Any two concepts in the ontology are linked in one of three different ways: they have no relation, one relation, or more than one relation. In our implementation, we used the Jena library for ontology reasoning to infer implicit relations. Given two documents d_1 and d_2, the set of relations between these documents are the combination of all extracted explicit and implicit relations between their respective classes, defined as followed:

$$R(d_1,d_2) = \cup_{x \in C_{d_1}, y \in C_{d_2}} [r_{explicit}(x,y) \cup r_{implicit}(x,y)] \quad (3)$$

where $r_{explicit}(x,y)$ and $r_{implicit}(x,y)$ denote the sets of explicit and implicit relations between each two concepts x and y, respectively. Also, let R_{xy} denote the set of all relations between x and y, including explicit and implicit ones:

$$R_{xy} = r_{explicit}(x,y) \cup r_{implicit}(x,y) \quad (4)$$

One of the significant steps to calculate the semantic relevance between two documents is to extract all explicit and implicit relationships between concepts based on an ontology. To illustrate, there are some concepts in the Sport ontology (introduced later) that have just a “is subclass of” relation as shown in Table 2, while other concepts have more than one relation as extracted and shown in Table 3.

Table 2: Examples of extracted explicit relations between different concepts in Sport ontology

Ontology module	Class name	Relation	Class name	Ontology module
sport.owl	InterdependentTeam	is subClass of	Team	sport.owl
sport.owl	GroupSystem	is subClass of	RankingSystem	sport.owl
sport.owl	SwissSystem	is subClass of	GroupSystem	sport.owl
sport.owl	Sport	is subClass of	Activity	sport.owl

Table 3: Examples of extracted explicit and implicit relations between two concepts in Sport ontology. The 1st row is an explicit relation while the 2nd row is an implicit relation.

Ontology module	Class name	Relation	Class name	Ontology module
sport.owl	Match	is subClass of	SportsEvent	sport.owl
sport.owl	Match	hasPart	SportsEvent	sport.owl

Semantic relevance metric By capturing the semantic relevancy, we can measure the relatedness between a group of concepts or a group of instances in the ontology. Given two concepts x and y belonging to documents d_1 and d_2, the semantic relevancy between these concepts is defined as follows:

$$sem_{rel}(x,y) = \frac{\sum_{r \in R_{xy}} w(r)}{Avg_{u \in C_{d_1}, v \in C_{d_2}} max(R_{uv})} \tag{5}$$

where $w(r)$ denotes the weight of a relation $r \in R_{xy}$ and $Avg_{u \in C_{d_1}, v \in C_{d_2}} max(R_{uv})$ is the maximum number of relations between any two concepts in the ontology. We assume that each relation has a weight value equal to 1. We define the document semantic relevancy by summing and averaging the semantic relevancies between their respective concepts, as defined as follows:

$$sem_{rel}(d_1,d_2) = \frac{\sum_{x \in C_{d_1}, y \in C_{d_2}} sem_{rel}(x,y)}{n_{d_1} \times n_{d_2}} \tag{6}$$

Where n_{d_1} and n_{d_2} are the total number of concepts in d_1 and d_2, respectively.

4 Datasets and Ontology

To evaluate our model, we need a dataset that contains real-world data leak and the associated ontology. We need a dataset, which can be categorized into a protected and a potentially leaked information from the dataset. Unfortunately, it is challenging to find such a dataset that is publicly available. However, we collected and organized a new DLP dataset, which can be used for DLP research. The ontology is a data model to represent a set of concepts and the relations between them in a specific domain of interest in a hierarchical structure. The ontology is used mainly for information representation, organization, and reasoning in a particular domain. Also, the concepts in an ontology may have explanation and definition to interpret concepts' usage in a particular ontology. In our research, we evaluate our model using an ontology and datasets related to sports domain of interest.

4.1 Sport Datasets

We evaluate our model in the sports domain using sport ontology and football datasets, which consist of BBC football news and Football leaks dataset [1] [15].

BBC football news dataset BBC sport news dataset consists of BBC sport news articles from 2004-2005. In addition, BBC sport news dataset is available publicly online in textual format for research purposes and covers 5 main sports fields including athletics, cricket, football, rugby, and tennis [1]. In our research, we used BBC football news dataset, which consists of 265 BBC football news articles and we consider them as non-sensitive files. Figure 7 shows a sample from the BBC football news dataset.

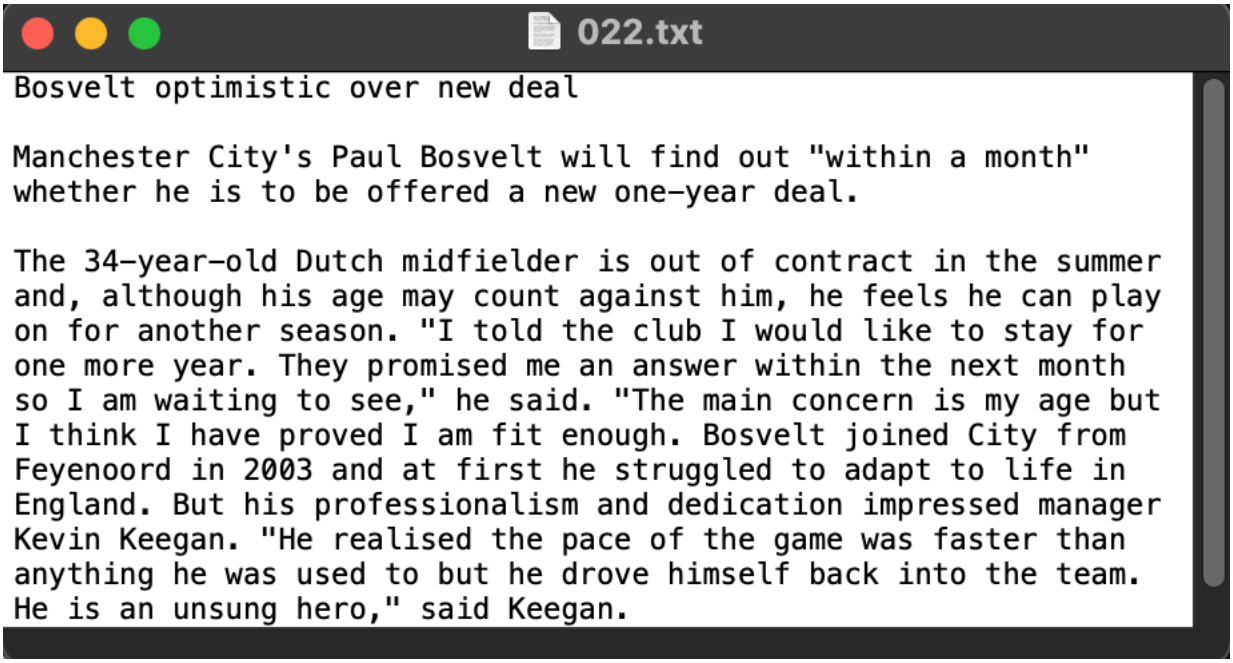

Figure 7: A sample from the BBC football news dataset

Football leaks dataset Football leaks dataset is based on one of the largest leaks in sports. In September 2015, football leaks occurred online on a website by Rui Pinto, who at the beginning went by the pseudonym, John. Later, it was found out that Rui Pinto was the actual person behind leaking sensitive football information about football clubs and famous footballers. Leaked data consisted of approximately 18.6 million documents such as emails, contracts, and spreadsheets [38] . Der Spiegel with the network European Investigative Collaborations (EIC) and its partners used these sensitive documents for investigation. This involved investigative reporting of confidential information by 80 journalists from 12 European media in 11 languages [38] [4]. The football leaks are considered a corruption of the European football industry while Rui Pinto declared that he just revealed illegitimate practices that happened within the football world [38]. In 2018, Buschmann and Wulzinger published a book on football leaks that uncovered the dirty deals behind football [15]. The book is structured into several short stories each related to different deals. We created a football leak dataset by saving the different stories in separate files. The football leak dataset consists of the collection of files generated from the aforementioned process. Figure 8 shows a sample text from the football leaks dataset.

The football leaks dataset has in total 203 files considered as sensitive files, which were grouped into two subsets: the first subset contains the first page of each leaked topic while the second subset has the remaining pages of each topic. Our goal is to use the first subset as the reference sensitive data, while the second subset is used for testing. So, in practice, the first subset will represent the protected data, while the second one will be used to test potential leak. Figure 9 shows the statistics of the football dataset

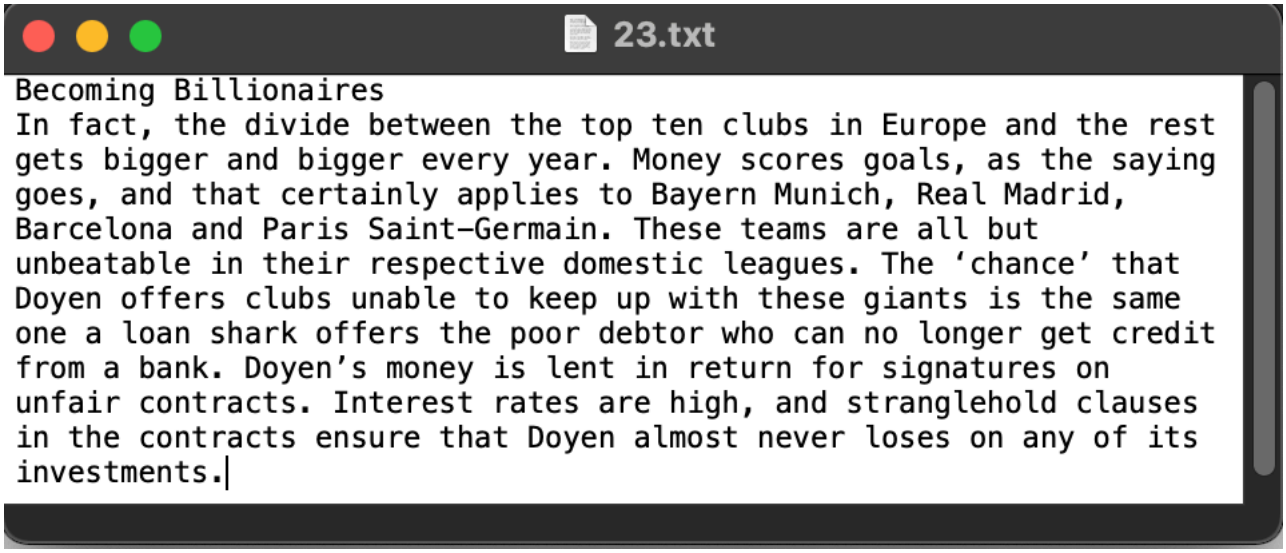

Figure 8: A sample text from the football leaks dataset

including football leaks and BBC football news.

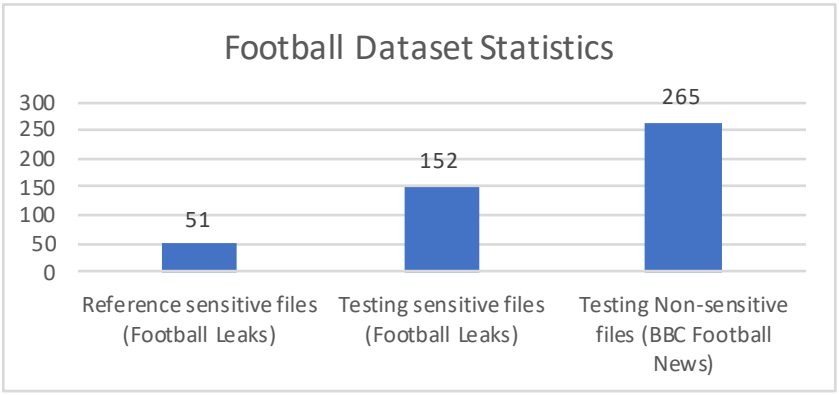

Figure 9: Football dataset statistics

4.2 Sport Ontology

We used Sport ontology, which covers football-related concepts and the relationships between them. Figure 10 below shows the concept tree of Sport ontology including concepts, relationships, and concepts' depth [6]. Also, Table 4 shows the Sport ontology's concepts, labels, and depths. Furthermore, a sport synonyms keywords' lexicon has been created and provided in our proposed model to assist in finding semantic keywords [7] [5].

5 Experimental Evaluation

In this section, we describe the experimental evaluation of our model and discuss the obtained results.

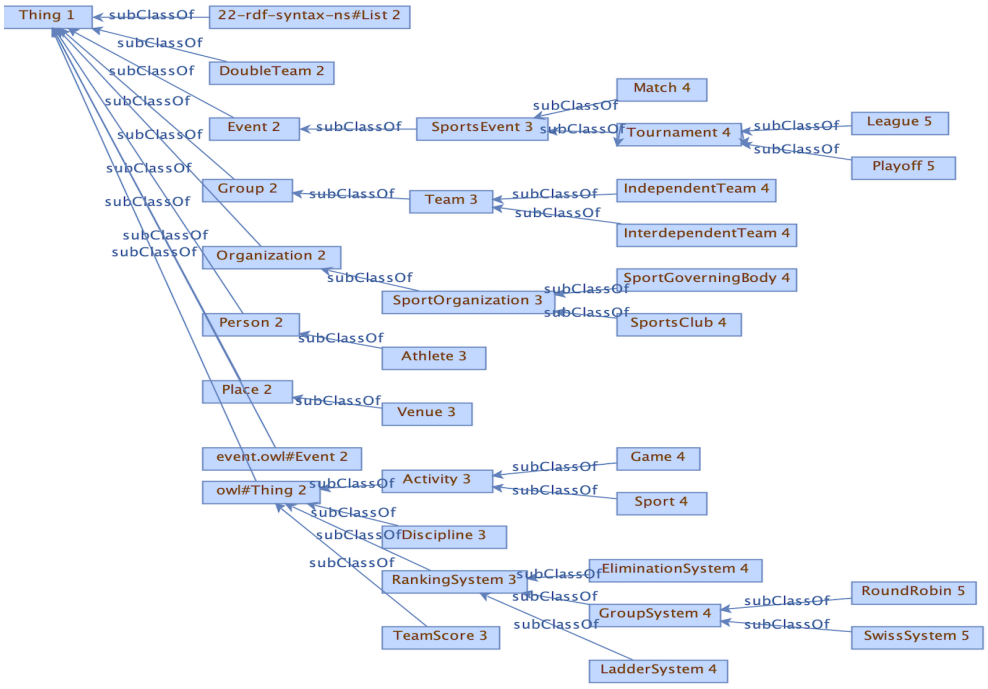

Figure 10: The concept tree of Sport ontology

Table 4: Sport Ontology concepts, labels, and depths

Label	Depth	Concept	Label	Depth	Concept
C_1	1	Thing	C_{18}	3	RankingSystem
C_2	2	22-rdf-syntax-ns#List	C_{19}	3	TeamScore
C_3	2	DoubleTeam	C_{20}	4	Match
C_4	2	Event	C_{21}	4	Tournament
C_5	2	Group	C_{22}	4	IndependentTeam
C_6	2	Organization	C_{23}	4	InterdependentTeam
C_7	2	Person	C_{24}	4	SportGoverningBody
C_8	2	Place	C_{25}	4	SportsClub
C_9	2	event.owl#Event	C_{26}	4	Game
C_{10}	2	owl#Thing	C_{27}	4	Sport
C_{11}	3	SportsEvent	C_{28}	4	EliminationSystem
C_{12}	3	Team	C_{29}	4	GroupSystem
C_{13}	3	SportOrganization	C_{30}	4	LadderSystem
C_{14}	3	Athlete	C_{31}	5	League
C_{15}	3	Venue	C_{32}	5	Playoff
C_{16}	3	Activity	C_{33}	5	RoundRobin
C_{17}	3	Discipline	C_{34}	5	SwissSystem

5.1 Evaluation Approach and Metrics

The evaluation of our model is based on the calculation of two performance metrics, which are the detection rate (DR) and the false positive rate (FPR). DR measures the ability of the model to detect data leakage, while FPR measures its ability to limit false alarms. A false positive occurs when a non-sensitive document is classified as sensitive. A false negative occurs when a sensitive document is falsely classified as non-sensitive; a true detection is just the opposite of a false negative, where a sensitive document is correctly classified as sensitive. The FPR and DR are defined as follows:

$$FPR = \frac{\text{Number of non sensitive documents classified as sensitive}}{\text{Total number of non sensitive documents}} \tag{7}$$

$$DR = \frac{\text{Number of sensitive documents classified as sensitive}}{\text{Total number of sensitive documents}} \tag{8}$$

We split the football leaked dataset into two subsets, datasets A and B as follows. Dataset A is considered as sensitive reference dataset, which consists of the first page of all topics while dataset B is considered as sensitive testing dataset, which consists of the remaining pages of all topics. In addition, the BBC football news dataset is considered as non-sensitive testing dataset and is called dataset C. We conducted the evaluation by applying two-fold cross validation on our model as follows. In the 1st round of two-fold cross validation, we check the similarity of each file in B against all files in A (used as reference) and calculate the number of files in B that are flagged as dissimilar as per our model (i.e. false negatives). Also, we check the similarity of each file in C against all files in A and calculate the number of files in C that have similarity in A (i.e. false positives) as shown in Figure 11. In the 2nd round of two-fold cross validation, we flip between datasets A and B. Then, we check the similarity of each file in A against all files in B and calculate the number of files in A that do not have similarity in B (i.e. false negatives). Also, we check the similarity of each file in C against all files in B and calculate the number of files in C that have similarity in B (i.e. false positives) as shown in Figure 11.

5.2 Evaluation Results

We conduct an experiment to apply the semantic similarity metric along with the semantic relevance metric on Sport domain including football datasets and Sport ontology. In this experiment, we aim to evaluate our model on the semantic similarity metric (2) combined with the semantic relevance metric (6) to measure the semantic overlap between documents and to detect suspicious documents. We conducted this experiment on sports domains of interests. We calculate the average DR and FPR based on two-fold cross validation. (Note: th_{sim} is the threshold of semantic similarity while th_{rel} is the threshold of relevance metric) Table 5 shows the performance results obtained by applying the combination of semantic similarity (2) and semantic relevance metrics (6) using the football dataset and sport ontology.

Table 5: Performance results obtained by applying the combination of semantic similarity and relevancy metrics on Football dataset.

		thr_{sim}= 0.65 AND thr_{rel} =0.5
Ontology modules		**Average**
sport	$DR(\%)$	72.79
	$FPR(\%)$	3.4

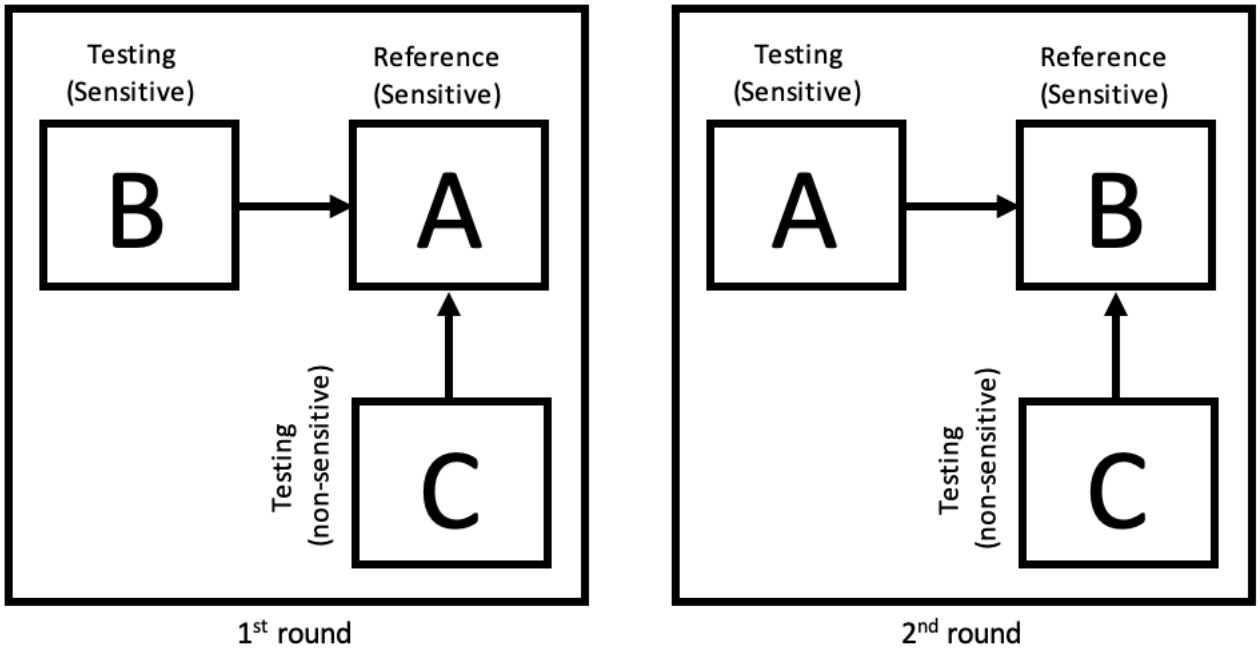

Figure 11: Two-fold cross-validation

The best results obtained using 0.65 for similarity threshold and 0.50 for relevancy threshold values were *DR*=72.79% and FPR=3.4%. Table 5 shows that we achieve satisfying performance by obtaining high *DR* and low FPR results of combining semantic similarity and semantic relevance metrics on sport domain. In other words, it can be noted that combining semantic similarity metric with semantic relevance metric gives better results in detecting semantic commonalities between any suspicious and sensitive documents.

6 Conclusion

In this paper, we are interested in detecting sensitive data loss or leakage out of organization's perimeter. We have proposed a new approach that focuses on detecting the commonalities between critical documents and suspicious documents, especially the documents that have been altered by rephrasing the text while the meaning is unchanged. This approach relies on information semantic including the semantic of concepts and their relations represented in an ontology. In addition, the matching of this approach is based on ontology based semantic metrics, which are semantic similarity metrics and semantic relevance metrics. Combining semantic relevance metric with semantic similarity metric produces encouraging performance results by generating relatively high detection rate and low false positive rate. Our current experiment was conducted by covering sports domain including two football datasets and sport ontology. In the future, we intend to run our model on different domains such as industrial, technical, medical domain. In addition, our goal is to build an ontology based on an organization's need and organization's critical data.

References

[1] Bbc datasets. http://mlg.ucd.ie/datasets/bbc.html",[Online; accessed on May 20, 2021].

[2] Data loss prevention, cisco. https://www.cisco.com/c/en/us/products/security/email-security-appliance/data-loss-prevention-dlp.html [Online; accessed on May 20, 2021].

[3] Data loss prevention, imperva. https://www.imperva.com/learn/data-security/data-loss-prevention-dlp/ [Online; accessed on May 20, 2021].

[4] De-spiegel. https://www.spiegel.de/international/world/interview-with-football-leaks-whistleblower-rui-pinto-a-1251121.htm [Online; accessed on May 20, 2021].

[5] Power thesaurus. https://www.powerthesaurus.org/sport/synonyms [Online; accessed on May 20, 2021].

[6] Sports ontology. https://github.com/Tobion/Sports-Ontology/blob/master/ontology.owl [Online; accessed on May 20, 2021].

[7] Thesaurus. https://www.thesaurus.com/browse/ [Online; accessed on May 20, 2021].

[8] Data loss prevention–keeping your sensitive data out of the public domain, 2011. https://slideplayer.com/slide/5286604/ [Online; accessed on May 20, 2021].

[9] Symantec data loss prevention solution., 2015. https://docs.broadcom.com/docs/data-loss-prevention-solution-en [Online; accessed on May 20, 2021].

[10] S. Alneyadi, E. Sithirasenan, and V. Muthukkumarasamy. A survey on data leakage prevention systems. *Journal of Network and Computer Applications*, 62:137–152, February 2016.

[11] W. J. Blanke. Data loss prevention using an ephemeral key. In *Proc. of the 2011 International Conference on High Performance Computing & Simulation (HPCSim'11), Istanbul, Turkey*, pages 412–418. IEEE, July 2011.

[12] W. Boehmer. Analyzing human behavior using case-based reasoning with the help of forensic questions. In *Proc. of the 24th IEEE International Conference on Advanced Information Networking and Applications (AINA'10), Perth, Western Australia, Australia*, pages 1189–1194. IEEE, April 2010.

[13] K. Borders and A. Prakash. Quantifying information leaks in outbound web traffic. In *Proc. of the 30th IEEE Symposium on Security and Privacy (S&P'09), Oakland, California, USA*, pages 129–140. IEEE, May 2009.

[14] I. Burdonov, A. Kosachev, and P. Iakovenko. Virtualization-based separation of privilege: Working with sensitive data in untrusted environment. In *Proc. of the First EuroSys Workshop on Virtualization Technology for Dependable Systems (VDTS'09), Nuremberg, Germany*, pages 1–6. ACM, March 2009.

[15] R. Buschmann and M. Wulzinger. *Football Leaks: Uncovering the Dirty Deals Behind the Beautiful Game*. Faber & Faber, 2018.

[16] M. Campbell. What are the consequences of data loss, 2021. https://www.unitrends.com/blog/what-are-the-consequences-of-data-loss [Online; accessed on May 20, 2021].

[17] V. R. Carvalho, R. Balasubramanyan, and W. W. Cohen. Information leaks and suggestions: A case study using mozilla thunderbird. In *Proc. of the 6th Conference on Email and Anti-Spam (CEAS'09), Mountain View, California, USA*, July 2009.

[18] P. Chahal, M. Singh, and S. Kumar. Relation based measuring of semantic similarity for web documents. *International Journal of Computer Applications*, 119(July), 2015.

[19] D. Clark, S. Hunt, and P. Malacaria. Quantitative analysis of the leakage of confidential data. *Electronic Notes in Theoretical Computer Science*, 59(3):238–251, November 2002.

[20] E. Costante, D. Fauri, S. Etalle, J. Den Hartog, and N. Zannone. A hybrid framework for data loss prevention and detection. In *Proc. of the 2016 IEEE Security and Privacy Workshops (SPW'16), San Jose, California, USA*, pages 324–333. IEEE, August 2016.

[21] E. Dosal. 8 data leakage prevention best practices, June 2018. https://www.compuquip.com/blog/8-data-leakage-prevention-best-practices [Online; accessed on May 20, 2021].

[22] D. Du, L. Yu, and R. R. Brooks. Semantic similarity detection for data leak prevention. In *Proc. of the 10th Annual Cyber and Information Security Research Conference (CISR'15), Oak Ridge, Tennessee, USA*, pages 1–6. ACM, April 2015.

[23] A. Gara. Morgan stanley fires rogue employee after customer data leak, January 2015. https://www.forbes.com/sites/antoinegara/2015/01/05/morgan-stanley-fires-rogue-employee-after-customer-data-leak/#39d2e8b67b7e [Online; accessed on May 20, 2021].

[24] J. M. Gómez-Hidalgo, J. M. Martín-Abreu, J. Nieves, I. Santos, F. Brezo, and P. G. Bringas. Data leak prevention through named entity recognition. In *Proc. of the IEEE Second International Conference on Social Computing (SocialCom'10), Minneapolis, Minnesota, USA*, pages 1129–1134. IEEE, August 2010.

[25] J. L. Griffin, T. Jaeger, R. Perez, R. Sailer, L. V. Doorn, and R. Cáceres. Trusted virtual domains: Toward secure distributed services. In *Proc. of the First IEEE Workshop on Hot Topics in System Dependability (HotDep'05), Yokohama, Japan*, pages 12–17. IEEE Press, June 2005.

[26] M. Hart, P. Manadhata, and R. Johnson. Text classification for data loss prevention. In *Proc. of the 11th International Symposium on Privacy Enhancing Technologies Symposium (PETS'11), Waterloo, Ontario, Canada*, volume 6794 of *Lecture Notes in Computer Science*, pages 18–37. Springer, Berlin, Heidelberg, July 2011.

[27] S. A. Kale and S. Kulkarni. Data leakage detection. *International Journal of Advanced Research in Computer and Communication Engineering*, 1(9):668–678, 2012.

[28] A. Kantor, L. Antebi, Y. Kirsch, and U. Bialik. Methods for document-to-template matching for data-leak prevention, 8 2012. US Patent 8,254,698.

[29] G. Katz, Y. Elovici, and B. Shapira. Coban: A context based model for data leakage prevention. *Information sciences*, 262:137–158, March 2014.

[30] K. Kaur, I. Gupta, and A. K. Singh. Data leakage prevention: E-mail protection via gateway. In *Proc. of the 10th International Conference on Computer and Electrical Engineering (ICCEE'17), Edmonton, Canada*, volume 933. IOP Publishing, May 2017.

[31] Y. Li, D. McLean, Z. A. Bandar, J. D. O'shea, and K. Crockett. Sentence similarity based on semantic nets and corpus statistics. *IEEE transactions on knowledge and data engineering*, 18(8):1138–1150, August 2006.

[32] H. Ling, Z. Chen, C. Yu, and Z. Cheng. Unified source user checking of tcp data packets for network data leakage prevention, July 2018. US Patent 10,015,145.

[33] H. Liu and P. Wang. Assessing text semantic similarity using ontology. *Journal of Software*, 9(2):490–497, February 2014.

[34] S. Liu and R. Kuhn. Data loss prevention. *IT professional*, 12(2):10–13, March 2010.

[35] A. Mahajan and S. Sharma. The malicious insiders threat in the cloud. *International Journal of Engineering Research and General Science*, 3(2):245–256, March-April 2015.

[36] J. Marecki, M. Srivatsa, and P. Varakantham. A decision theoretic approach to data leakage prevention. In *Proc. of the IEEE Second International Conference on Social Computing (SocialCom'10), Minneapolis, Minnesota, USA*, pages 776–784. IEEE, August 2010.

[37] R. Mogull. Understanding and selecting a data loss prevention solution. https://securosis.com/blog/understanding-and-selecting-a-data-loss-prevention-dlpcmfcmp-solution-part- [Online; accessed on May 20, 2021], September 2007.

[38] R. Pinto. Football leaks, 2019. https://footballleaks2015.wordpress.com [Online; accessed on May 20, 2021].

[39] S. Reading. Data loss prevention: A holistic approach, dlp whitepaper, 2016. https://securereading.com/wp-content/uploads/2016/08/Data-Leakage-Prevention.pdf [Online; accessed on May 20, 2021].

[40] S. Saad and I. Traore. Heterogeneous multi-sensor ids alerts aggregation using semantic analysis. *Journal of Information Assurance and Security*, 7(2):79–88, 2012.

[41] L. Securosis. Understanding and selecting a data loss prevention solution, 2017. https://zbook.org/read/9c625_understanding-and-selecting-a-data-loss-prevention-solution.html [Online; accessed on May 20, 2021].

[42] Y. Shapira, B. Shapira, and A. Shabtai. Content-based data leakage detection using extended fingerprinting. *arXiv preprint arXiv:1302.2028*, 2013.

[43] X. Shu and D. D. Yao. Data leak detection as a service. In *Proc. of the 8th International Conference on*

Security and Privacy in Communication Systems (SecureComm'13), Padua, Italy, pages 222–240. Springer, Berlin, Heidelberg, September 2012.

[44] M. Sokolova, K. El Emam, S. Rose, S. Chowdhury, E. Neri, E. Jonker, and L. Peyton. Personal health information leak prevention in heterogeneous texts. In *Proc. of the 2009 workshop on adaptation of language resources and technology to new domains (AdaptLRTtoND'09), Borovets, Bulgaria,*, pages 58–69. ACM, September 2009.

[45] A. Squicciarini, S. Sundareswaran, and D. Lin. Preventing information leakage from indexing in the cloud. In *Proc. of the IEEE 3rd International Conference on Cloud Computing (CLOUD'10), Miami, Florida, USA*, pages 188–195. IEEE, 2010.

[46] T. A. Staff. The biggest data breaches, April 2019. `https://www.techworld.com/security/uks-most-infamous-data-breaches-3604586/` [Online; accessed on May 20, 2021].

[47] R. Tahboub and Y. Saleh. Data leakage/loss prevention systems (dlp). In *Proc. of the World Congress on Computer Applications and Information Systems (WCCAIS'14), Hammamet, Tunisia*, pages 1–6. IEEE, January 2014.

[48] T. Wüchner and A. Pretschner. Data loss prevention based on data-driven usage control. In *Proc. of the IEEE 23rd International Symposium on Software Reliability Engineering (ISSRE'12), Dallas, Texas, USA*, pages 151–160. IEEE, 2012.

[49] S. Yoshihama, T. Mishina, and T. Matsumoto. Web-based data leakage prevention. In *Short Paper Proc. of the 5th International Workshop on Security (IWSEC'10), Kobe, Japan*, pages 78–93. Information Processing Society of Japan, November 2010.

Author Biography

Hanan Waheed Alhindi[1] is an assistant professor in the Information Technology Department at King Saud University where she has been a faculty member since 2010. She is the instructor of several undergraduate topics in cybersecurity, computer networks, and programming languages. Hanan completed her Ph.D. in Computer Engineering at the University of Victoria, Victoria, BC., Canada. Also, she finished her master's degree in Computer Engineering at King Saud University, Riyadh, Saudi Arabia and she is considered the first female Saudi student, who has received a master degree in Computer Engineering from King Saud University. She has completed her undergraduate studies in Information Technology at King Saud University, Riyadh, Saudi Arabia. Hanan's research interests lie in the area of computer and network security. She published several papers in scientific journals and conferences. In addition, she has received the best paper award of the 2nd International Conference on Wireless, Intelligent, and Distributed Environment for Communication (WIDECOM 2019), Milan, Italy.

[1] No photo is available



Issa Traore obtained a PhD in Software Engineering in 1998 from Institute Nationale Polytechnique (INPT)-LAAS/CNRS, Toulouse, France. He has been with the faculty of the Department of Electrical and Computer Engineering of the University of Victoria since 1999. He is currently an Associate Professor and the Coordinator of the Information Security and object Technology (ISOT) Lab (http://www.isot.ece.uvic.ca) at the University of Victoria. His research interests include biometrics technologies, computer intrusion detec- tion, network forensics, software security, and software quality engineering. He has published over 100 technical papers in computer security in the last 10 years. He has served as Associate Editor for the International Journal of Communication Networks and Distributed Systems (IJCNDS).



Isaac Woungang received the Ph.D. degree in mathematics from the University of the South, Toulon-Var, France, in 1994.From 1999 to 2002, he was a Software Engineer 928 with Nortel Networks, Ottawa, ON, Canada. Since 2002, he has been with Ryerson University, Toronto, 930 ON, where he is currently a Professor of computer 931 science. He has authored or coauthored eight books 932 and more than 90 refereed technical articles in scholarly international journals and proceedings of international conferences. His current research interests include radio resource management in next-generation wireless networks, big 936 data, Internet of Things, and cloud computing. Dr. Woungang has Guest Edited several Special Issues with various reputed 938 journals such as Computer Communications (Elsevier) and Telecommunication 939 Systems (Springer). He was the Chair of Computer Chapter, IEEE Toronto Section, from 2012 to 2018.