Data Leak Prevention through Named Entity Recognition

José María Gómez-Hidalgo[†], José Miguel Martín-Abreu[†],
Javier Nieves[§], Igor Santos[§], Felix Brezo[§], and Pablo G. Bringas[§]

[†]Optenet
Madrid, Spain
Email: {jgomez, jabreu}@optenet.com

[§]S^{3}Lab, DeustoTech, University of Deusto
Bilbao, Spain
Email: {javier.nieves, isantos, felix.brezo, pablo.garcia.bringas}@deusto.es

***Abstract*—The rise of the social web has brought a series of privacy concerns and threats. In particular, data leakage is a risk that affects the privacy of not only companies but individuals. Although there are tools that can prevent data losses, they require a prior step that involves the sensitive data to be properly identified. In this paper, we propose a new automatic approach that applies Named Entity Recognition (NER) to prevent data leaks. We conduct an empirical study with real-world data and show that this NER-based approach can enhance the prevention of data losses. In addition, we present and detail the implementation of a prototype built with these techniques and show how it can be used by both particulars and companies in order to handle data losses.**

I. Introduction

The use of the Internet has been growing over the last few years both by individuals and companies. The so-called *social web* is one of the reasons underlying this success. Specifically, Emerging Interaction Environments (EIE) such as Facebook[1], Blogger[2], LinkedIn[3], or Second Life[4] are used by millions of users and have become a commercial and collaboration tool for companies.

However, as any successful media, EIEs are prone to misuse and, consequently, they are being used for malicious behaviours including spam and malware distribution [1], [2], sexual harassment, bullying, and brand or individual image damaging. In particular, there can also be a risk concerning the privacy of individuals or the sensitive data of the companies because of the high number of possible interactions [3], [4], [5], [6].

Data Leak Prevention helps to stop the loss of sensitive data [7]. Most of these tools are very effective when protecting already known private information. Nevertheless, a great amount of private information is not recognised until it has been disclosed to unknown users or competition enterprises.

Given this background, we propose here the first system for data leak prevent that uses Named Entity Recognition (NER). NER is a sub-area within information extraction that intends to locate and classify elements in text into known categories such as the names, organisations, places and so on [8]. These techniques have been applied in biomedical text [9], identification of genes and proteins [10], improvement of web search [11], or spell correction [12].

Specifically we make the following contributions:

- We show how to use NER to prevent data leaks.
- We provide an empirical study of NER techniques using a real-world dataset in two different languages: English and Spanish.
- We present a prototype able to alert users about possible data leaks.

The remainder of this paper is organised as follows. Section II introduces NER techniques and how are they utilised in this research. Section III details the performed experiment and evaluates obtained results whereas section IV details the implementation of a prototype as the application of this paper. Finally, section V concludes and outlines the avenues of the future work.

II. Background on Named Entity Recognition

Named Entity Recognition (NER), also called *entity identification and entity extraction*, is a subtask of information retrieval that intends to identify and classify atomic elements which appear in a text into predefined categories [13].

The NER systems use both linguistic grammar-based techniques and statistical models. Although the first type usually obtains better accuracy, it requires a hard work developed by experienced computational linguists. Otherwise, the second type of NER techniques does not present this issue, however, it needs a large amount of manually annotated training data. Nonetheless, these statistical models can be quickly re-adapted to extract different types of entities without annotated training data because they rely on unsupervised or semi-supervised learning.

[1] http://www.facebook.com/

[2] http://www.blogger.com

[3] http://www.linkedin.com/

[4] http://secondlife.com/

Currently, the NER techniques are widely used into two domains: (1) the journalistic field and (2) biological field. Regarding the journalistic scope, these techniques usually try to extract entities like names of people, places and organisations. In the biological field, NER techniques extract entities such as genes, proteins, diseases and drugs [14]. In this research, we deal with texts closer to the journalistic field than the biological one.

The main objective of the current researches in this area is to utilise NER techniques that do not depend on the language, in other words, multilingual NER techniques without reliance on any specific domain. These type of NER techniques arise a constraint that make us use the same method in many different languages. However, the extension to new languages should be quick and simple. This raises two choices:

- Using supervised learning approaches based on attributes of the language which are independent on the representation. It is strongly recommended the existence of manually annotated text collections for each target language to achieve a high accuracy level.
- Using unsupervised learning techniques or clustering, which do not depend on the aforementioned collections but whose effectiveness is often less than in the previous method.

In this work, we concentrate on the first group of algorithms because there must be a minimum efficiency level to make further applications feasible.

The best way to make a comparative analysis between NER techniques is to gather information from scientific competitions. For instance, in *CoNLL Shared Task*, multiple developed researches were compared by international research teams using standard and common libraries. Both editions of this competition were relevant, the first one, in 2002[5], worked on Spanish and Dutchman languages, while the second one, in 2003[6], worked on English and German. In both of them the main purpose was to detect entities extracted from texts that had been published on news. Fig. 1 shows one example of the developed work.

```
[PER Wolff ] , currently a journalist in [LOC
Argentina ] , played with [PER Del Bosque ] in
the final years of the seventies in [ORG Real
Madrid ].
```

Fig. 1. Categorised Text Example.

In Fig. 1, entities marked in brackets are identified using NER techniques and and each of them has associated a type, being $PER = person$, $ORG = organization$ and $LOC = location = place$.

The majority of the developed systems used by different research groups in both competitions accomplished their training using supervised-learning techniques, applied to characterise data in terms of the following types of attributes:

- Spelling elements and capital letters, dots which identify potential candidates for named entities (i.e., names).
- Low complex syntactic elements, such as syntactic categories (i.e., noun, verb, adverb) associated with each word.
- Lexical frequency elements, which is the frequency of certain words in the vicinity of a proper name, computed on a statistically way.

Regarding the classification engine, these systems have been implemented employing many learning algorithms, including Artificial Neural Networks (ANN), naive Bayesian classifiers, learning decision trees, learning classification rules, the Support Vector Machines (SVM) among others. In general, if the data representation is adequate and there is a large amount of data recorded, the selection of a learning algorithms is one of the less relevant parameters on the accuracy of the systems.

III. Experimental Study

A. Generation of the Dataset

We extracted a dataset in order to test the validity of our method. We selected Twitter[7] as the source for the dataset for the following reasons:

- Twitter offers a huge amount of data: everyday hundreds of millions of comments from 23 millions of users are stored (according to Quancast[8]).
- Commonly, Twitter's comments express different opinions about news, services and brands. Therefore, they are an appropriate source of data for the experimentation with NER techniques.
- The data within Twitter is visible for any user. In fact, tools exist that allow to search and open-source programming APIS (e.g., JTwitter[9]).

To extract the dataset, we used Google Insights[10] that can compare patterns in the search traffic in certain regions, categories, time intervals or properties. This tool permitted us to list the most popular search terms associated to a particular category of products. The terms were utilised to select products, brands, services or individuals, to be employed in turn in order to find possible related comments. We used the most popular categories [15] (refer to Table I).

For each of these categories, we obtained the most popular searches from the months of January to August in 2009. The results were manually examined to determine which of the searches refer to products, services, brands, people or places. Then, each of the searches was conducted on Twitter, following the next specifications:

- We reserved the results of the first 3 searched comments or *tweets*[11] for training, and the results of the last 2 for testing or evaluation.

[5] http://www.cnts.ua.ac.be/conll2002/ner

[6] http://www.cnts.ua.ac.be/conll2003/ner

[7] http://twitter.com/

[8] www.quantcast.com/

[9] http://www.winterwell.com/software/jtwitter.php

[10] www.google.com/insights/search/

[11] In Twitter jargon, a tweet refers to a user's comment

TABLE I
POPULAR CATEGORIES.

CATEGORY	SUB-CATEGORIES
Books & Literature	N/A
Automotive	N/A
Food & Beverage	N/A
Shopping	Shoes T-shirts Fashion Designers Underwear and Lingerie Watches and Complements Retail Clothing Stores
Finances & Assurances	N/A
Video & Photography	N/A
Electronic and Informatics	Consumer electronics Hardware Computer Security Software
Games	N/A
Leisure and Entertainment	Music Movies
Telecommunications	Mobile Phones Wireless Devices
Travelling	Travel Flights

- In order to experiment with English and Spanish, so the techniques used are sufficiently robust to changing the language, we compiled comments throughout Twazzup[12] multilingual search engine, and we also searched tweets in each language separately.

Within each sub-dataset (training and test datasets written in Spanish and English), we removed the comments in other languages because, commonly, Twazzup retrieved several results written in other languages.

TABLE II
THE 88 PERFORMED SEARCHES IN TWITTER.

acer aspire	harry potter	nike	sony vaio
adidas	hawaii	nikon	star wars
adobe acrobat	hollister clothing	nintendo	stephenie meyer
airfare	honda	nokia	tesco
amazon	hp	north face	the north face
asus	ibanez	norton	timex
banana republic	iphone	olympus	toyota
bbq	ipod	omega	transformers
blackberry	jordan	panasonic	turbotax
bmw	kaspersky	ps3	twilight
bose	kitchenaid	psp	v neck
canon	kodak	puma	vans
citizen	las vegas	quicken	vero moda
converse	lego	ralph lauren	vibram five fingers
dell	les paul	rimowa	victoria secret
ds	lumix	rims	vintage clothing
ebay	marshalls	rolex	vizio
ed hardy	mcafee	samsung	wheels
fender	microsoft	seiko	wii
garmin	microsoft office	shakespeare	windows xp
gibson	michael jackson	sony	xbox
hanes	new york	sony ericsson	xp

In this way, we compiled a maximum of 80 comments for each of the sought entities (products, services and so on). It is worth to mention that repetitions occurred (for example, 'nike' appeared in both purchasing and clothing), and we did not reached to the target number of five relevant searches in certain categories. The final number of obtained tweets for each sub-collection was 4240 for training and 2800 for testing (for both English and Spanish).

Table II shows the final 88 performed searches to obtain the comments. Several brands (e.g., 'lego', 'converse'), products (e.g., 'ipod', 'wii'), people (e.g, 'Michael Jackson', 'Stephenie Meyer'), places (e.g., 'New York', 'Las Vegas') appeared . It should be noticed that we considered online stores as brands for all purposes (e.g., 'v neck', 'bbq').

B. Analysis

On the basis of the aforementioned dataset, we used the *Freeling*[13] natural language processing package and performed the following steps:

1) We selected 4 categories from the most popular ones: shoes, consumer electronic, books & literature, and computer security.
2) From the available *tweets*, we removed for each tested language (English and Spanish) the ones that were not written in it. We obtained a total number of 170 comments in Spanish and 76 in English.
3) Then, we applied a tokenisation step (divide the content of the text into words), obtaining 3880 words or tokens in Spanish and 1654 in English. Note that Freeling's tokeniser is not able to split artificially-composed words (e.g., 'nikerules') or divided using underlines (e.g., 'Nike_Air').
4) Finally, the comments were processed with Freeling, using the NER module available for each language.

TABLE III
CONFUSION TABLE OF THE SPANISH DATASET.

SPANISH CORPUS	NAMED ENTITY	NOT NAMED ENTITY
Named Entity	702	338
Not Named Entity	14	2826

TABLE IV
CONFUSION TABLE OF THE ENGLISH DATASET.

ENGLISH CORPUS	NAMED ENTITY	NOT NAMED ENTITY
Named Entity	156	60
Not Named Entity	67	1371

Table III shows the obtained confusion matrix for the Spanish corpus whereas Table IV shows the results for the English dataset. The results for the main diagonal correspond to correct classifications, while the another diagonal shows the errors. In terms of overall efficiency, we obtained a success rate of 90.92% for Spanish and 92.32% for English.

The errors located in the upper right corner of the matrices are false positives, namely, expressions misidentified as Named Entities by the system. It is important to keep false positives low, because, the lower the number, the fewer times the system alerts the user about an entity that really is not. Otherwise, false negatives (i.e., expressions which are entities but have not

[12] http://www.twazzup.com/

[13] http://www.lsi.upc.edu/∼nlp/freeling

been identified as such by the system) located in the lower left corner are possible sources of leaks not detected by the system and unnoticed for the user, making them more dangerous.

In terms of false positives, regardless we obtained quite high results, the errors may be re-iterative, in other words, the system alerts an user several times about the same entity (which actually is not). This situation can be handled through a *white list* where the false positives are stored so they will not be considered in the future.

Regarding false negatives, Spanish dataset only missed detection of the 2% of the entities while, surprisingly, the test with the English corpus nearly the 30% were not detected. Notwithstanding, when we examined the produced errors, most of them corresponded with entities actually detected in other occurrences.

IV. Prototype

After passing the previous stages, we developed a prototype that includes the obtained results. The experimental prototype implemented a data leak prevention function employing NER techniques supported by Freeling. This prototype is able to alert the administrator about the apparition of suspicious entities. Therefore, the administrator can make decisions about what to do in each case (e.g, block, allow them to pass but register them, and ignore them).

The aim of this prototype is to demonstrate the feasibility of integrating NER techniques positively evaluated in the previous steps. In a real environment, the prototype must be deployed like *Security as a Service* (SaaS). Nevertheless, we conducted the experiment on an *endpoint* or PC because its development is considerably easier for this platform.

A. Functional Definition

This prototype is able to detect candidates of sensitive data and propose possible solutions to the administrator. Currently, we had already developed a data leak prevention software, which also includes features for Web filtering, mail filtering, antivirus, personal firewall, and antiphishing. The functions of data leakage prevention are included in the section of phishing, within the area of protection of personal data.

The personal data protection permits us to enter, modify and delete personal data such as credit cards, passwords, account numbers, phone numbers, addresses and other personal data in order to ensure that this information is used in a safe and not fraudulent way when we are surfing the Internet. Fig. 2 shows the management screen for the protection of personal data.

In the configuration screen, the protection of personal data can be activated or deactivated as well as the action to be accomplished when the program detects an attempt to send personal data stored on the computer. The possible actions are the following.

- **Mask:** When a user enters any personal data in a Website, this data is sent in an encrypted manner so that the recipient can not recognise them.
- **Warning:** When a user inserts personal data in a Website, it will display a message showing that the use is trying to send personal information and will ask the user to enter an administrator password. If the password is correct, the data will be sent normally. Otherwise (or if you press the 'Cancel' button), information is sent encrypted.

The 'Report' checkbox lets the user specify whether to send an email to the stored address each time it is detected an attempt to send personal data. In order to configure the personal data, the user introduces his name (or nickname), his credit card number, password, telephone number, address and personal password and press the 'Add' button. To be more reliable, some values are asked again.

To delete personal data, the user selects the item he wants to delete in the right box and clicks on 'Delete' button. To modify a value, it is selected in the right box and user clicks on 'Edit'. The values appear into the fields to the left, in which they can be modified. If there are changes in the name (or nickname), a new record will be added.

In order to inspect the data sent to the Internet when the patterns defined by user are detected, the most simple and appropriate way to include proactive detection of new patterns are the following:

1) Each time the user makes an outgoing connection, the system inspects and applies the NER techniques to detect new expressions needing protection.
2) If the system finds a pattern, it notifies to the user in order to take the decision to store it or not.
3) Provided that the user chooses to add this new pattern, it is included in the same list in which manually-defined patterns are stored.
4) Even if the user does not chooses to add the new pattern, this pattern will be stored in a separate list so that if in the future is detected again, it will not be shown to the user.
5) The patterns have an expiration date, which allows to ask to the user about discovered patterns after a certain time. This process ensures that the adopted patterns are always updated.

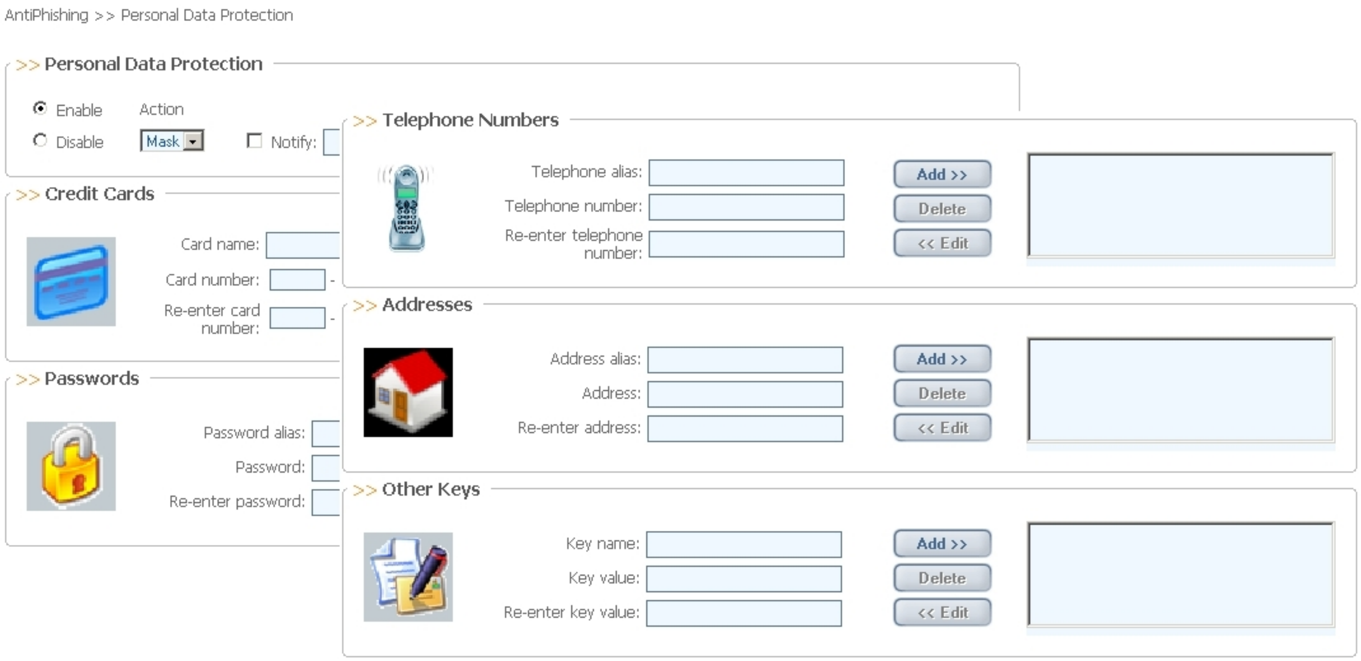
The prototype works like a personal firewall, which is trained or adapted by the user along the time. At first, it is more intrusive, however, later, it is much less or nothing intrusive when the patterns are stabilised.

B. Design and Implementation

To be minimally intrusive on the existing architecture and promote the integration with previously developed algorithms, the solution was developed as independent modules which can communicate with each other.

We introduced a *hook* in a existing system, which writes into hard drive all outbound connections. Every information is stored through temporary files in a shared folder. The NER system operates independently with the files as an ongoing process performing the following tasks:

1) Examining the first file available.



AntiPhishing >> Personal Data Protection

>> Personal Data Protection

● Enable ○ Disable

Action: Mask

□ Notify:

Fig. 2. The management Web for configuring the protection of personal data.

2) Applying the NER system.
3) If an entity is detected, it is added to the encrypted configuration file of user's patterns.
4) The file is erased and we go to the first step.

This mode of work is not totally adjusted to behaviour defined in the previous subsection, however, it allows fast coding and debugging. At the contrary, the operational mode mentioned before is simulated by manual inspection of the list of patterns in the configuration of the protection of personal data. When the user finds one that was not manually entered, it is actually an entity suggested by the system. If the pattern does not interest to the user, he can remove it using the interface and manually add it to the exceptions file so that the system does not reinsert it into the list of patterns.

V. Conclusions and Future Work

One of the most dangerous current security threats for social networks users is private information leaks. From phishing attacks to unnoticed users mistakes, leaks are getting more and more common, and data leak prevention tools must be made much more effective in order to preserve user and corporate privacy. One way to solve these problems is to apply *text classification techniques* that are used to process any text disclosed to a social network to search for named entities like people, organisation or product names. In this paper, we use a technique called Named Entity Recognition (NER). Systems based on NER techniques must be prepared to work with every language and they must not be dependent on a specific domain.

We have extended the state of the art to be capable of avoiding the data leakage: (i) developing a new corpus based on social networks' comments in two languages (Spanish and English), (ii) we have tested the NER techniques learnt with our corpus, and finally, (iii) we have developed a prototype to prove that it is possible to detect data leakage of both individuals and companies. The experimental evaluation arises the following conclusions:

1) The current NER techniques based on spelling, contextual and lexical attributes are effective on the western languages, and allow good accuracy achieved with homogeneous sets of attributes to deal with general-purpose entities (people, organizations, brands, location).
2) From the operational point of view, the errors in the classifications are almost always solvable.

In summary, the results are very promising, specially, when systems implement and integrate NER techniques in environments with data leaks prevention.

Indeed, the future development of this prototype is oriented in two main ways. Firstly, we plan to incorporate this prototype to a commercial solution. Finally, we plan to perform more experiments to tune the system allowing to achieve better accuracy level.

Acknowledgment

This research was financially supported by the Ministry of Industry of Spain, project Cenit SEGUR@, Security and Trust in the Information Society, (BOE 35, 09/02/2007, CDTI).

References

[1] Z. Mazur, H. Mazur, and T. Mendyk-Krajewska, "Security of Internet Transactions," *Internet-Technical Development and Applications*, p. 243, 2009.

[2] W. Luo, J. Liu, J. Liu, and C. Fan, "An analysis of security in social networks," *Dependable, Autonomic and Secure Computing, IEEE International Symposium on*, vol. 0, pp. 648–651, 2009.

[3] E. Zheleva and L. Getoor, "To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles," in *Proceedings of the 18^{th} international conference on World wide web*. ACM New York, NY, USA, 2009, pp. 531–540.

[4] B. Krishnamurthy and C. Wills, "On the leakage of personally identifiable information via online social networks," in *Proceedings of the 2nd ACM workshop on Online social networks*. ACM, 2009, pp. 7–12.

[5] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, "Inferring private information using social network data," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 1145–1146.

[6] B. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala, "Privacy-Preserving Data Publishing," *Foundations and Trends in Databases*, vol. 2, no. 1-2, pp. 1–167, 2009.

[7] I. Abbadi and M. Alawneh, "Preventing insider information leakage for enterprises," in *Emerging Security Information, Systems and Technologies, 2008. SECURWARE'08. Second International Conference on*, 2008, pp. 99–106.

[8] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Named Entities: Recognition, classification and use*, p. 3, 2009.

[9] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii, "Tuning support vector machines for biomedical named entity recognition," in *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*. Association for Computational Linguistics, 2002, p. 8.

[10] L. Tanabe, N. Xie, L. Thom, W. Matten, and W. Wilbur, "GENETAG: a tagged corpus for gene/protein named entity recognition," *BMC bioinformatics*, vol. 6, no. Suppl 1, p. S3, 2005.

[11] M. Pasca, "Acquisition of categorized named entities for web search," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM, 2004, pp. 137–145.

[12] P. Ruch, R. Baud, and A. Geissb
"uhler, "Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record," *Artificial intelligence in medicine*, vol. 29, no. 1-2, pp. 169–184, 2003.

[13] E. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," *Development*, vol. 922, p. 1341, 1837.

[14] A. Cohen and W. Hersh, "A survey of current work in biomedical text mining," *Briefings in Bioinformatics*, vol. 6, no. 1, p. 57, 2005.

[15] B. Jansen and A. Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Information Processing & Management*, vol. 42, no. 1, pp. 248–263, 2006.