Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 7038–7043
Marseille, 11–16 May 2020

HITZALMED: Anonymisation of Clinical Text in Spanish

Salvador Lima, Naiara Perez, Laura García-Sardiña, Montse Cuadros
SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Donostia/San-Sebastián, 20009, Spain
{slima, nperez, lgarcias, mcuadros}@vicomtech.org

Abstract

HITZALMED is a web-framed tool that performs automatic detection of sensitive information in clinical texts using machine learning algorithms reported to be competitive for the task. Moreover, once sensitive information is detected, different anonymisation techniques are implemented that are configurable by the user –for instance, substitution, where sensitive items are replaced by same category text in an effort to generate a new document that looks as natural as the original one. The tool is able to get data from different document formats and outputs downloadable anonymised data. This paper presents the anonymisation and substitution technology and the demonstrator which is publicly available at https://snlt.vicomtech.org/hitzalmed.

Keywords: clinical data, sensitive data detection, anonymisation

1. Introduction

Data has become an invaluable resource for both research and commercial purposes. When it comes to data containing personal information, such as health records, there is an ethical and legal responsibility towards respecting the individuals' privacy. This has led to the introduction of specific laws that address this issue, such as the European Union's General Data Protection Regulation (GDPR) directive or the United States' Health Insurance Portability and Accountability Act (HIPAA).

A possible solution to the privacy problem is anonymisation. Anonymisation can be defined as the process of removing all information from a document that could potentially point to a given person, such as names, phone numbers or e-mail addresses. The resulting documents allow us to use real data that cannot be linked to a real person. Medlock (2006) describes three different approaches to anonymisation: *removal* ('replacing a reference with a blank place-holder'), *categorisation* ('replacing a reference with a label in some way representing its type or category') and *pseudonymisation* ('replacing a reference with a variant of the same type').

Manual anonymisation is a tiresome and expensive process. For this reason, considerable efforts are being made to automatise the task. This is straightforward in the case of structured text or tabular data, but the task becomes considerably more challenging when dealing with unstructured natural language.

This article presents HITZALMED[1], a web-framed tool that assists with the anonymisation of clinical free text in Spanish. HITZALMED uses a hybrid approach that combines Machine Learning (ML) techniques to detect Protected Health Information (PHI) and a more traditional rule-based system for their de-identification. The main features of HITZALMED are presented below:

- It supports multiple document formats
- It features an automatic PHI recogniser and classifier that has proven to be competitive in the *MEDDOCAN: Medical Document Anonymization* (Marimon et al., 2019) shared task, having achieved F1-scores higher than 0.95
- The classifier distinguishes 21 fine-grain PHI categories (e.g., patient vs doctor name)
- HITZALMED comes with two strategies for PHI anonymisation: categorisation and pseudonymisation
- The PHI recognition, classification, and anonymisation proposed by HITZALMED can be easily edited, corrected or new ones can be added using the web interface

The structure of this article is as follows: Section 2. provides a brief overview of related work, mostly focusing on sensitive information detection and anonymisation techniques and tools; Section 3. introduces HITZALMED, both the user interface and its capabilities (Section 3.1.), and HITZALMED's inner workings (Section 3.2.); Section 4. concludes the paper and introduces future lines of work.

2. Related Work

Multiple automatic anonymisation systems have been proposed over the years. These systems have to first detect and classify any personal information and then treat it using different techniques.

For the former task, oftentimes these systems use two methodologies: either a pattern-matching approach or Machine Learning (ML). One of the earliest systems is Scrub (Sweeney, 1996), released in 1996 for Electronic Health Records in English. Scrub uses algorithms based on rules and dictionaries to detect categories such as names, addresses, cities or countries. In a similar fashion to other Natural Language Processing (NLP) tasks, ML methods dominate most of the recent publications due to their efficiency. However, datasets containing sensitive information can be hard to come across. For this reason, shared tasks have played an important role in furthering research by publicly releasing annotated corpora. Specifically, the i2b2 de-identification challenges (Uzuner et al., 2007; Stubbs et al.,

[1] https://snlt.vicomtech.org/hitzalmed

2015) have gathered a lot of interest and its two corpora are widely used in research. Some authors, such as Dernoncourt et al. (2016) or Khin et al. (2018), using Deep Learning architectures, achieve 0.9783 and 0.9787 F-1 score, respectively, at PHI detection on the i2b2 dataset. Finally, although not as common, some authors also use Information Theory techniques for this task (Sánchez et al., 2013).

After detection, the second part of the anonymisation process is sanitising sensitive information. Several freely-available tools exist that redact sensitive information from free text in English. One instance is MIST (MITRE Identification Scrubber Toolkit) (Aberdeen et al., 2010), which, in a similar fashion to HitzalMed, first annotates the target phrases and then proceeds to replace them either by categorisation or pseudonymisation. Other similar tools include LingPipe (Carpenter, 2007), MITdeid (Neamatullah et al., 2008) or NLM-Scrubber (Kayaalp et al., 2013).

Languages other than English are also seeing some developments in this direction. Some examples include Mamede et al. (2016), who built a complete system with detection and anonymisation for Portuguese, or Tveit et al. (2004), who developed a semi-automatic anonymisation system for clinical texts in Norwegian.

In Spanish, some recent studies include Medina and Turmo (2018), who present a Spanish-Catalan Health Records corpus annotated with PHI; Hassan et al. (2018), who describe a detection method using named entities; and García-Sardiña (2018), who presents an annotated, anonymised corpus of spontaneous dialogue data and Knowledge Transfer techniques for sensitive data identification. Additionally, in 2019 the first community challenge about anonymisation of medical documents in Spanish, MEDDOCAN[2] (Marimon et al., 2019), was held as part of the IberLEF initiative. Its authors studied the GDPR and i2b2 specifications and released a synthetic corpus of 1,000 clinical studies enriched with PHI. The challenge included two different tasks that would be evaluated on the said corpus: *a)* NER offset and entity type classification, and *b)* sensitive span detection.

There are also some automatic tools specifically designed for both detection and sanitation of text in Spanish. For instance, Iglesias et al. (2008) present MOSTAS, a morpho-semantic tagger, anonymiser and spell-checker system for biomedical texts that returns documents annotated in XML format. Additionally, several enterprises offer commercial solutions for anonymisation of legal documents[3,4], but we did not find any publications describing these systems. As far as we know, HitzalMed is the only publicly available web-demonstrator for text anonymization in Spanish.

3. HitzalMed

HitzalMed is an environment that enables the detection of sensitive data by applying machine learning algorithms, their substitution with anonymised data, and the edition of the resulting detected and substituted data.

HitzalMed has a web front-end with a simple and clean interface that can perform three main steps as shown in Figure 1. First, it processes the input document provided by the user. HitzalMed supports a wide range of document formats (e.g., TXT, DOCX, PDF) as it uses tika-python[5] to perform text extraction. Secondly, it recognises and classifies PHI in the given document. Thirdly, the recognised PHI are automatically substituted to generate an anonymised version of the input document, which can then be manually checked and edited by the user. This final revision is of great importance in any automated anonymisation process and must be carried out to ensure that no linkable personal data persist in the resulting text.

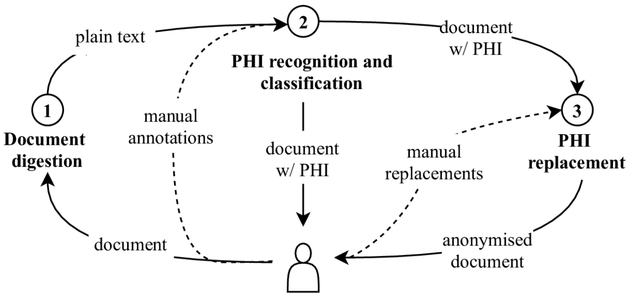


Figure 1: HitzalMed workflow

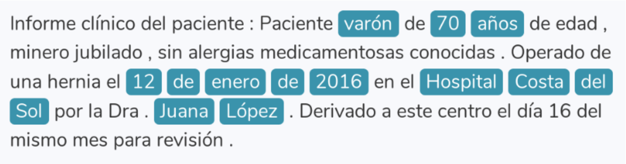
The following sections describe firstly the web interface to understand the workflow and secondly the technical details.

3.1. Interface

After a short registration process, users can choose to learn about the technical side of our PHI detection system –by downloading our scripts and models, which are freely available– or they can use the web interface to upload their own documents and see the different anonymisation techniques in action.

An overview of the website's layout can be seen in Figure 2. HitzalMed offers users the option to upload documents in multiple text formats and treat them at four different levels. We will now go through them using a sample text to show how every step applies to the same document.

1. **Identification:** All detected items are highlighted, allowing users to take a quick glance at the big picture of the sensitive items in the text. Figure 3 shows this step. Users can make corrections to the automatic detection (i.e., add or remove annotations) by double-clicking on the affected token. For instance, in the example we are using, 'minero' (*miner*) has not been detected as a profession; we can annotate and classify the word manually, as shown in Figure 4.



Informe clínico del paciente : Paciente varón de 70 años de edad , minero jubilado , sin alergias medicamentosas conocidas . Operado de una hernia el 12 de enero de 2016 en el Hospital Costa del Sol por la Dra . Juana López . Derivado a este centro el día 16 del mismo mes para revisión .

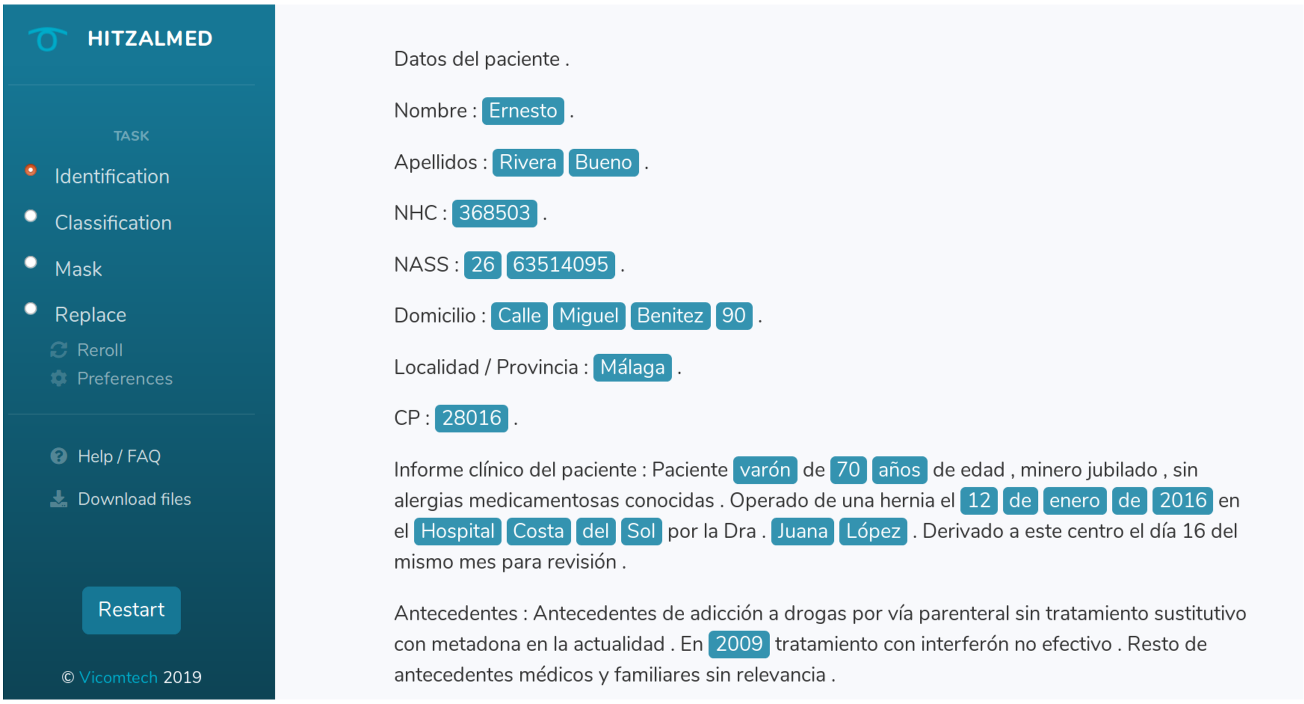
Figure 3: Identification in HitzalMed

[2] http://temu.bsc.es/meddocan/

[3] https://www.serikat.es/eJusticia

[4] https://www.garrigues.com/latam/commitment/innovacion

[5] https://github.com/chrismattmann/tika-python



Datos del paciente .

Nombre : Ernesto .

Apellidos : Rivera Bueno .

NHC : 368503 .

NASS : 26 63514095 .

Domicilio : Calle Miguel Benitez 90 .

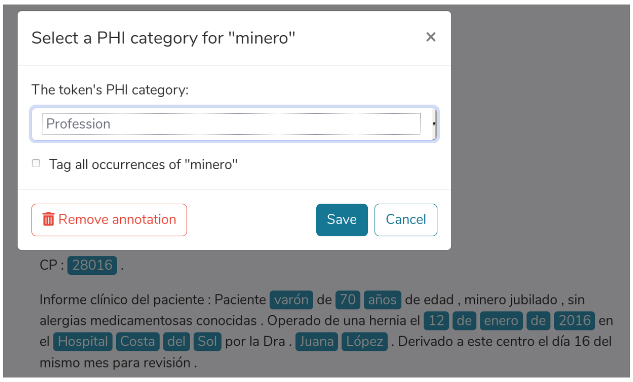
Localidad / Provincia : Málaga .

CP : 28016 .

Informe clínico del paciente : Paciente varón de 70 años de edad , minero jubilado , sin alergias medicamentosas conocidas . Operado de una hernia el 12 de enero de 2016 en el Hospital Costa del Sol por la Dra . Juana López . Derivado a este centro el día 16 del mismo mes para revisión .

Antecedentes : Antecedentes de adicción a drogas por vía parenteral sin tratamiento sustitutivo con metadona en la actualidad . En 2009 tratamiento con interferón no efectivo . Resto de antecedentes médicos y familiares sin relevancia .

Figure 2: HITZALMED overview



The token's PHI category: Profession

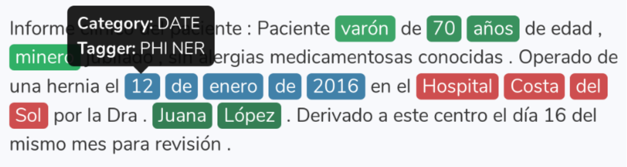
□ Tag all occurrences of "minero"

CP : 28016 .

Informe clínico del paciente : Paciente varón de 70 años de edad , minero jubilado , sin alergias medicamentosas conocidas . Operado de una hernia el 12 de enero de 2016 en el Hospital Costa del Sol por la Dra . Juana López . Derivado a este centro el día 16 del mismo mes para revisión .

Figure 4: Tagging a profession that had not been detected

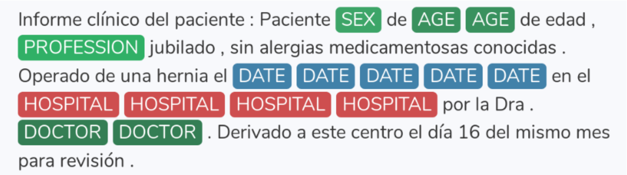
2. **Classification:** Each PHI is classified into one of 21 different categories, introduced in Table 1. Each category is highlighted in a different colour, so that they can be easily distinguished at first glance. Figure 5 shows how our text from the previous examples would look in this mode.



Inform nte : Paciente varón de 70 años de edad , miner ergias medicamentosas conocidas . Operado de una hernia el 12 de enero de 2016 en el Hospital Costa del Sol por la Dra . Juana López . Derivado a este centro el día 16 del mismo mes para revisión .

Figure 5: Classification in HITZALMED

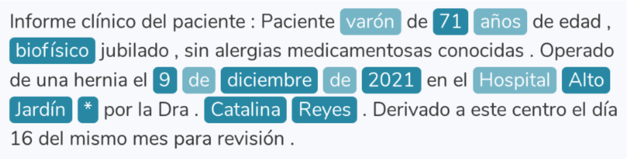
3. **Masking:** This is one of the two types of anonymisation that HITZALMED offers. Masking is akin to categorisation, one of the approaches by (Medlock, 2006) described in the introduction. Each item is simply replaced with its category, as shown in Figure 6.



Informe clínico del paciente : Paciente SEX de AGE AGE de edad , PROFESSION jubilado , sin alergias medicamentosas conocidas . Operado de una hernia el DATE DATE DATE DATE DATE en el HOSPITAL HOSPITAL HOSPITAL HOSPITAL por la Dra . DOCTOR DOCTOR . Derivado a este centro el día 16 del mismo mes para revisión .

Figure 6: Masking in HITZALMED

4. **Replacement:** The other anonymisation technique offered by HITZALMED. A new version of the text is generated in which sensitive items have been substituted by a randomly-generated item of the same category (see Figure 7). The result of this approach preserves readability and is much more natural-looking than using Masking. Users are able to process their documents even further in this mode. For example, they can choose to change the proposed replacements by *rerolling*. This will re-run the process, showing a new result every time. Figure 8 shows an example of a different possible output obtained by rerolling. It is also possible to manually edit any substitution by double-clicking on it. Furthermore, users can decide how much dates will shift by choosing a range of days, months, and years through the *Preferences* menu.



Informe clínico del paciente : Paciente varón de 71 años de edad , biofísico jubilado , sin alergias medicamentosas conocidas . Operado de una hernia el 9 de diciembre de 2021 en el Hospital Alto Jardín * por la Dra . Catalina Reyes . Derivado a este centro el día 16 del mismo mes para revisión .

Figure 7: Replacement in HITZALMED

ID Numbers	People	Dates	Locations	Contact Info	Other
Doctor D	Doctor	Date	Street	E-mail address	Other
Patient ID	Patient		Healthcare centre	Phone number	
Insurance ID	Age		Hospital	Fax number	
Contact ID	Sex		Institution		
	Profession		Location		
	Kinship		Country		

Table 1: Sensitive information categories recognised by HITZALMED

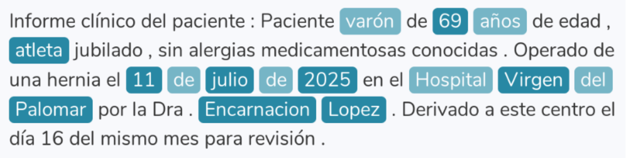

Informe clínico del paciente : Paciente varón de 69 años de edad , atleta jubilado , sin alergias medicamentosas conocidas . Operado de una hernia el 11 de julio de 2025 en el Hospital Virgen del Palomar por la Dra . Encarnacion Lopez . Derivado a este centro el día 16 del mismo mes para revisión .

Figure 8: Different replacement of the same text after rerolling

Finally, clicking on the download button offers the option of downloading the original and the replaced documents in TXT format, and the different set of tags serialized with pickle.

3.2. Technical Description

As presented at the beginning of this section, HITZALMED automatically recognises and classifies sensitive data and proposes substitutes that are editable by the user.

3.2.1. PHI Recognition and Classification

HITZALMED automatically recognises and classifies PHI from the extracted plain text using a model based on spaCy's[6] EntityRecognizer module and trained on data from the MEDDOCAN challenge (Table 2).

	train	dev	test
# document	500	250	250
# tokens	260,407	138,812	132,961
vocabulary	26,355	15,985	15,397
# PHI	11,333	5,801	5,661

Table 2: Size of the MEDDOCAN corpus

spaCy's entity recogniser is built on Bloom Embeddings (Serrà and Karatzoglou, 2017) and residual Convolutional Neural Networks (He et al., 2016). We followed the given recipe [7] with default settings and applied the recommended tweaks: compounding batch size, dropout decay, and parameter averaging.

The resulting model can recognise a total number of 21 different PHI categories, which are listed in Table 1. These classes were defined by the MEDDOCAN shared task organisers, who adapted the HIPAA guidelines by adding new categories (such as the patient's age) and removing others that did not fit within the Spanish healthcare system (e.g., health plan beneficiary numbers).

The official results obtained at the shared task with this model are shown in Table 3. This model obtained the 20th position among the 63 participating systems in the challenge. For detailed information on how the model was learned, we refer the reader to the dedicated article (Perez et al., 2019).

task	P	R	F1
PHI recognition	0.967	0.953	0.960
+ classification	0.965	0.948	0.956

Table 3: HITZALMED's PHI recognition and classification results at the MEDDOCAN challenge

3.2.2. Anonymisation by Substitution

Whereas masking –anonymisation by categorisation, the first technique offered by HITZALMED– is as simple as replacing each sensitive item with its tag, substitution is substantially harder as we need to find replacements that are both natural and adequate for the target items in the document. We use three different approaches depending on the complexity of the tag, namely, regular expressions (RegEx), dictionaries, and a combination of the two. These approaches are presented below.

It must be noted that there are two PHI categories that are never automatically pseudonymised. The first one is the **Sex** tag. We considered that even if we replaced gender-specific words (such as 'varón', *boy*, and 'mujer', *woman*) with more gender-neutral options, Spanish morphological features would still allow the reader to infer the sex of the patient. Moreover, the sex of the patient might be in close relation with the events described in the document.

The second one is the **Other** tag, which encompasses many dissimilar types of information, making it extremely hard to find an appropriate replacement automatically. As users are able to edit any item manually, we leave these special cases' substitution for their consideration.

Finally, an important feature of this mode is that replacements are consistent within a same document; that is, replacements will be reused if a tagged item appears more than once in the text. Date items are also altered by the same combination of days, months and years so that temporal coherence is kept throughout the document.

RegEx-based Methods

Regular expressions are used to find the relevant spans to be substituted within the detected PHI. For example, given

[6] https://spacy.io

[7] https://spacy.io/usage/training#ner

the tagged phrase '20 years of age' we must locate '20' as the part that needs to be replaced.

Identifiers and other numbers Regular expressions are used here to detect numeric expressions within PHI. The tags **Doctor's ID**, **Patient's ID**, **Insurance's ID**, **Contact's ID**, **Telephone**, and **Fax number** are made up of a series of numbers of a fixed length, which are simply replaced by a random series of numbers of the same length.

Age For this tag, we use regular expressions to locate the numbers in the detected PHI. These are randomly changed within a default interval of [-3, +3] (customisable by the user through the interface), unless the age is under 14 years old, as we reckoned that this kind of information may be meaningful in some contexts.

E-mail address We also use regular expressions to make sure that phrases automatically tagged as e-mail addresses are actually so; then, they are simply replaced with the generic string 'nombre.apellido@anon.com' (*name.surname@anon.com*) in order to avoid generating an existing address.

Dictionary-based Methods

For some tags, the simplest solution was to select a random replacement from a dictionary that includes items of the same category.

Country These items are simply replaced by another country's name randomly chosen from a dictionary of almost two hundred country names.

Profession In the case of this tag, we first analyse the morphological features of the tagged word; then, a random profession is retrieved from a hand-crafted list that contains over 200 profession names conjugated for both genders.

Kinship A list of all possible family relations was created and used as source for the replacements. It was divided into four different lists depending on the gender of the words as well as whether they describe a relationship of descendance (*younger than*) or ascendance (*older than*). This was done to avoid generating awkward sentences by replacing, for example, 'grandmother' with 'granddaughter'. Before choosing a replacement, we check to which of the four lists the original word belongs. A random member is then taken from the corresponding list.

Names and surnames We grouped together **Doctor** and **Patient** tags, as both are used for person names and surnames. Three censuses were gathered from the Instituto Nacional de Estadística (INE; *Spanish Statistical Office*): one for male names, another for female names and a last one for surnames. Each contains over 25,000 items ordered by frequency. Replacements for each token in a phrase tagged as **Doctor** or **Patient** are picked from the census the token belongs to. If a token is found in more than one census, the list in which the name is relatively most frequent is chosen to draw the replacement. In the rare scenario that a token is not in any of the censuses, a random name is drawn from a gender-neutral list computed from the intersection of the male and female censuses. The replacements are picked from the 100 most common items in each census in order to generate more generic outputs, which we view as a desirable outcome for anonymisation.

Mixed Methods

For some classes, in order to create a replacement that is faithful to the original string, we need to divide inputs into smaller parts. For this, we combine both regular expressions and word lists.

Location The items of this class may be either city names, ZIP codes or contain both at the same time. For that reason, we use a simple regular expression to separate letters from numbers. Again, numbers are replaced with another random number of the same length; letters are replaced by a random city drawn from a list that contains over sixty different Spanish city names.

Street The complexity of this tag lies in the rich variety of different elements it can contain. Addresses may include street numbers, door numbers or letters, stairs, and so on. We try to detect any of these items with regular expressions, randomise them, and attach the result to a random combination of a road type (e.g. 'calle', *street*, 'avenida', *avenue*, and so on) and street name. Both of these are taken from lists. The dictionaries of road types and street names were computed from a gazetteer of addresses provided by the organisers of the MEDDOCAN challenge[8].

Healthcare Facilities The tags **Healthcare Centre**, **Hospital** and **Institution** are similar, so they were treated equally. First, we created fictional names for the main types of health facilities: hospitals, clinics, institutes, residences, and healthcare centres. Using regular expressions, we try to recognise healthcare facilities' terms in order to maintain casing and spelling variations (e.g., 'Hospital', 'H.' , 'hptal', and so on). If we find any, we classify it into one of the healthcare facility types mentioned before. If no match is found, the original item is replaced by a random healthcare facility type and a random name from our list.

Date The date expressions present in a document are parsed into full dates by means of heuristics. Then, the same number of days, months, and years are subtracted or added to all the dates, in order to preserve the original timeline described in the document. The number of days, months, and years to add or subtract is chosen randomly for each document from a given interval –by default, 1 to 31 days, 1 to 12 months, and 1 to 10 years; the user can modify these ranges as desired from the interface. Each new date is finally converted to the same format as the respective original date expression. This means, for instance, that the new date will use the same separators (e.g. slashes,

[8] http://temu.bsc.es/meddocan/index.php/resources/

dashes, ...), or that if the date expressions contain a month name instead of the corresponding number, the substitution will do so as well.

4. Conclusions

This article has introduced HITZALMED, a publicly available web-framed tool for the automatic anonymisation of Spanish clinical text. HITZALMED is based on three main steps. Firstly, it processes a document from a wide range of document formats to convert it into plain text. Secondly, it recognises automatically personal sensitive information (PHI) with a proven competitive machine learning algorithm and, finally, it substitutes PHI to obtain an anonymised output file. As future work, we plan to evaluate the performance of the substitution methods, and assess how valid real users find the final anonymisation output. Additionally, we plan to convert the web-framed tool into a tool which could be installed on-premises.

5. Acknowledgements

This work has been supported by Vicomtech and partially funded by the project DeepReading (RTI2018-096846-B-C21, MCIU/AEI/FEDER,UE).

6. Bibliographical References

Aberdeen, J., Bayer, S., Yeniterzi, R., Wellner, B., Clark, C., Hanauer, D., Malin, B., and Hirschman, L. (2010). The MITRE Identification Scrubber Toolkit: Design, training, and assessment. *International Journal of Medical Informatics*, 79:849–59.

Carpenter, B. (2007). LingPipe for 99.99% recall of gene mentions. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 307–309.

Dernoncourt, F., Lee, J. Y., Uzuner, Ö., and Szolovits, P. (2016). De-identification of Patient Notes with Recurrent Neural Networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

García-Sardiña, L. (2018). Automating the anonymisation of textual corpora. Master's thesis, University of the Basque Country (UPV/EHU).

Hassan, F., Domingo-Ferrer, J., and Soria-Comas, J. (2018). Anonimización de datos no estructurados a través del reconocimiento de entidades nominadas. In *Actas de la XV Reunión Española sobre Criptología y Seguridad de la Información (RECSI 2018)*, pages 102–106.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Iglesias, A., Castro, E., Pérez, R., Castaño, L., Martínez, P., Gómez-Pérez, J. M., Kohler, S., and Melero, R. (2008). MOSTAS: Un Etiquetador Mrfo-semántico, Anonimizador y Corrector de Historiales Clínicos. *Procesamiento del lenguaje Natural*, 41:299–300.

Kayaalp, M., Browne, A., Callaghan, F., Dodd, Z., Divita, G., Ozturk, S., and Mcdonald, C. (2013). The pattern of name tokens in narrative clinical text and a comparison of five systems for redacting them. *Journal of the American Medical Informatics Association*, 21:423–431.

Khin, K., Burckhardt, P., and Padman, R. (2018). A Deep Learning Architecture for De-identification of Patient Notes: Implementation and Evaluation. *arXiv*, abs/1810.01570.

Mamede, N., Baptista, J., and Dias, F. (2016). Automated anonymization of text documents. In *Proceedings of the 2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 1287–1294.

Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodríguez, H., Lopez Martin, J. A., Villegas, M., and Krallinger, M. (2019). Automatic De-Identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, pages 618–638.

Medina, S. and Turmo, J. (2018). Building a Spanish/Catalan Health Records Corpus with Very Sparse Protected Information Labelled. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Medlock, B. (2006). An Introduction to NLP-based Textual Anonymisation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

Neamatullah, I., Douglass, M. M., Li-wei, H. L., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., and Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1).

Perez, N., García-Sardiña, L., Serras, M., and Del Pozo, A. (2019). Vicomtech at MEDDOCAN: Medical Document Anonymization. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, pages 696–703.

Serrà, J. and Karatzoglou, A. (2017). Getting Deep Recommenders Fit: Bloom Embeddings for Sparse Binary Input/Output Networks. In *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys 2017)*, pages 279–287.

Stubbs, A., Kotfila, C., and Uzuner, O. (2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of Biomedical Informatics*, 58(Suppl):S11–S19.

Sweeney, L. (1996). Replacing Personally-Identifying Information in Medical Records, the Scrub System. *Proceedings of the AMIA Fall Symposium*, pages 333–337.

Sánchez, D., Batet, M., and Viejo, A. (2013). Automatic General-Purpose Sanitization of Textual Documents. *IEEE Transactions on Information Forensics and Security*, 8(6):853–862.

Tveit, A., Edsberg, O., Røst, T., Faxvaag, A., Nytrø, Ø., Nordgård, T., Ranang, M., and Grimsmo, A. (2004). Anonymization of General Practioner Medical Records. In *Proceedings of the Second HelsIT Conference*.

Uzuner, O., Luo, Y., and Szolovits, P. (2007). Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association*, 14:550–63, 06.