BERT and Approximate String Matching for Automatic Recognition and Normalization of Professions in Spanish Medical Documents

Víctor Suárez-Paniagua[1,3] and Arlene Casey[2]

[1] Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom.
[2] Advanced Care Research Centre, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom.
[3] Health Data Research UK, London, United Kingdom.
{vsuarez,arlene.casey}@ed.ac.uk

Abstract. This publication presents the participation of the EdIE-KnowLab team in the MEDical DOcuments PROFessions recognition shared task from IberLeF 2021. The proposed system consists of a Spanish version of the BERT classification model, BETO, for the Named Entity Recognition tasks and an approximate string matching technique using Damerau–Levenshtein distance for the Normalization task. The NER systems reached 64.3% and 60.4% in Micro-Average F1 for Task 1 and Task 2, respectively. The approximate string matching approach obtained 17.8% in F1 for the Normalization task. Source code to reproduce the results is available under the MIT license at https://github.com/vsuarezpaniagua/EdIE-MEDDOPROF.

Keywords: Named Entity Recognition · Normalization · Medical Documents · Deep Learning · BERT · Damerau–Levenshtein.

1 Introduction

Determining occupational status of a patient is important information that can be used to model and provide surveillance about disease and has the potential to provide intervention strategies. Recent work has demonstrated the importance of understanding occupational status and how it links to aspects of health e.g., health inequality [29], occurrence of chronic disease [36], and mental health [3]. Understanding patient occupation has become more acutely relevant with the recent Covid-19 pandemic where specific occupation roles have been more adversely impacted than others. Occupational information though is often found in free-text as opposed to structured fields in medical texts. Chilmane et al. [8] show that compared to structured fields, where only 14% of records hold occupational

IberLEF 2021, September 2021, Málaga, Spain.

status, using NLP on the free-text records of patients they can increase patient record occupational status to 57%. Extracting information like occupation status from free-text is known as Named Entity Recognition (NER). NER tasks seek to locate and classify named entities in unstructured text into pre-defined categories, such as person names, locations, medical codes, time expressions, occupations. The challenge with free-text medical records is they are highly unstructured, non-standardised, and can lack semantic and syntactic cohesiveness [1].

This work describes our submissions to the MEDical DOcuments PROFessions recognition shared task (MEDDOPROF) [24] which is part of the IberLEF2021. In this task participants must develop an automated approach to detect occupation or occupational status, who the occupation is referring to and finally the normalization to a valid occupational code within Spanish medical documents. Our main approach is to use BETO [7], a BERT pre-trained model for Spanish NER tasks, and an approximate string matching technique using Damerau–Levenshtein distance for the Normalization task.

2 Related Work

The NER task within the medical field is not new with early research based on systems, such as MedLEE [17], MetaMap [2] and cTAKES [31]. Most of these early systems though are rule-based and supported with medical lexicon approaches. In more recent times the task of NER has been approached using pre-trained language models [35], particularly Bidirectional Encoder Representations from Transformers (BERT) [10]. This approach is outperforming other methods [23,11] and this trend continues in the biomedical domain [22,18].

Occupation status has been researched within the framework of de-identification, detection of personal information in order to remove or replace, anonymising the text [25,6]. Ahmed et al. [1] benchmark NER tasks on clinical text for de-identification. They compare a variety of deep learning methods and show Bi-LSTM CRF to be overall best, outperforming Transformer methods alone. However, looking at the individual entities the Transformer method provides the highest Recall for *Profession*. In addition, they comment that clinical texts suffer from varying lengths and believe this impacted the Transformers ability to adequately detect long term dependencies but that the model could be further improved by expanding the Transformer architecture beyond two layers of multi-level attention.

In recent years there have been shared tasks which focus on Spanish health or medical text and include extraction of occupation status. For example, MEDDOCAN [26] use Spanish medical text for NER detection in a de-identification task, which included *Profession* as an entity. The task was based on a synthetic corpus of clinical documents, with many well performing entries based on LSTM CRF architectures [30,33,12]. Lange et al. [21] the top performing entry with a Bi-LSTM model investigated several embedding types as input. Whilst performance differences were small they concluded that domain-specific

input representations perform best for Recall and the domain-independent input representations perform best in terms of Precision.

In ProfNer [28], the best results are obtained by the teams that have used Transformer-based architectures. The ProfNER task focuses on the detection of profession and occupational status on Spanish health related social media text. The top ranking entry in this task [5] used a BETO classifier with Hyperparameter Optimization (HPO) to fine-tune the model parameters. They compare two systems, a Transformer based model and a RNN model. As the performance is better in BETO they align this to the transfer capabilities of using a pre-trained language model and the ability to fine-tune to the task. They do not use any external knowledge, such as an occupation dictionary, but believe this could further improve performance. Other participants show that inclusion of a custom dictionary improves performance [27]. Yaseen et al. [37] demonstrate that the stacking of different embeddings improves the overall score. Specifically for NER they use Spanish BERT, fastText [4], BytePair sub-word [19] but do not fine tune any of the embeddings.

Our main approach is to use a Transformer architecture with the Spanish BERT approach, BETO, for the NER tasks. The model combines multiple BETO classifiers that are fine-tuned independently for each entity type. We did not utilise any dictionary knowledge or additional embeddings but these may be interesting angles to explore in the future to improve performance.

3 Dataset

The MEDDOPROF corpus [15,14] contains 1500 and 344 Spanish clinical cases for the training set and the test set from 20 different specialties. A team of clinical experts and linguists annotated the medical document following a guideline [16] using Brat Standoff format [32]. Task 1 consists of the classification of occupation mentions such as a profession (*Profesion*), an employment status (*Situacion Laboral*), and an activity (*Actividad*). Task 2 involves the recognition of the person who the occupation is referring to such as the patient (*Paciente*), a health profession (*Sanitario*), a family member (*Familiar*), or to someone else (*Otros*). Task 3 determines the normalization of each mention to one of the valid codes from a list [13] extracted from SNOMED-CT and the European multilingual classification of Skills, Competences and Occupations (ESCO). Table 1 shows the total number of mentions for each entity type in the Task 1 and Task 2.

The organizers also provided a complementary entity dataset of additional labelled mentions that can be used by participants with clinical entities such as symptoms, diseases, drugs and procedures, and linguistic entities such as negation trigger, uncertainty trigger and their scopes. However, in our experiments we decided to use only the information given by the training set.

3.1 Data preprocessing

The first stage taken is to prepare the data for the NER classifier. The medical documents were transformed into lower case, some special characters were

Table 1. Number of instances for each class in the training and development sets from the highest represented to the lowest.

Entity type	training set	test set
Profesion	2528	566
Situacion Laboral	1011	85
Actividad	119	16
Paciente	1735	592
Sanitario	1231	294
Familiar	207	53
Otros	485	146

replaced by a white space and the sentences were tokenized using the Spanish pipeline of spaCy [20]. In addition, the annotations were marked with the BIOES encoding, which is an extension of the BIO tag schema [34], where the tags 'B', 'I', 'E', 'S' indicate the token positions in the mentions as the beginning, the inside, the ending, and the single token entity, respectively, and the 'O' represents the tokens that are not entities.

In the corpus there are some clinical cases that contain more than four thousand tokens. However, the current BETO classifier only allows sentences with 512 tokens or less. For this reason, the medical documents were split into different parts with the maximum number of sentences until they reach 512 tokens or lower. The span of these split parts are saved to keep track of the offsets after recognizing the entities for the final prediction.

Some annotated documents contain embedded entities, these annotations were deoverlapped by taking the longest mention and filtering out any additional entity contained within it. For instance, the sentence '*En Diciembre inicia un negocio con pareja.*' ('*In December he/she starts a business with a partner.*') has two annotated mentions '*inicia un negocio*' ('*start a business*') and '*inicia un negocio con pareja*' ('*start a business with a partner*') where the first entity was discarded and the second entity is kept as it is the longest one. Moreover, we used multiple NER classifiers, one per each entity type, to solve the problem of the overlapping mentions with different classes.

4 Methods

The annotations in the medical documents were divided for each entity type after the preprocessing phase. For Task 1 and Task 2, a single uncased BETO NER classifier was applied for each type of mention to be trained and validated independently in order to obtain the best performance. Figure 1 shows the single BETO classifiers applied to a medical document for each class and, then, the predictions were merged to create the final annotation.

During model development we observed that some entity types were not widely represented within the training set and the BETO model learned to

recognize all the tokens as 'O'-tag. To solve this problem, a basic undersampling technique was introduced, which filters out all the documents that do not contain any positive tag ('B','I','E', or 'S'). Thus, forcing the BETO classifier to learn from these examples instead of having many documents without entities. We refer to this approach as BETO+Positive in the experiments while BETO+ALL refers to the approach that uses all the training set. In addition, we incorporated training and validation sets from the ProfNER Task [28] in order to have a greater representation of professions. These models are denoted as BETO+ProfNER in the results.

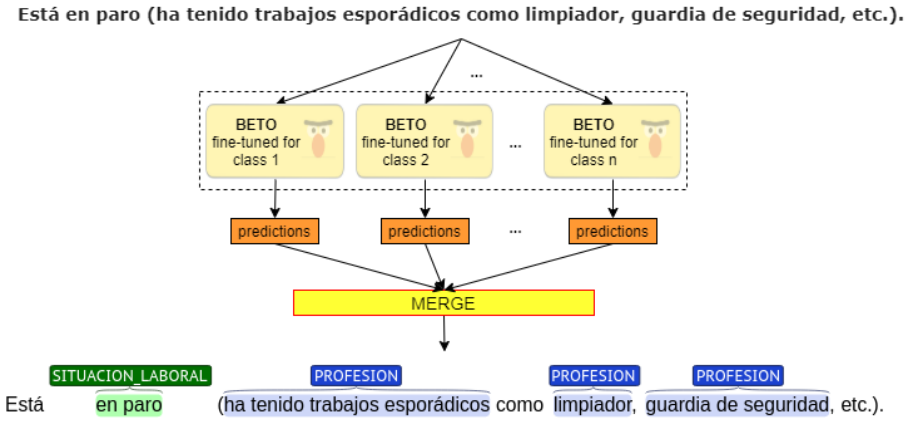

Fig. 1. The proposed NER system that recognizes and classifies the mentions in a given Spanish radiology report.

For Task 3, the normalization of the professions, we compared each recognized entity to the provided list of valid codes using an exact string matching technique. Then, we applied an approximate string matching technique called Damerau–Levenshtein distance [9] for those mentions that were not assigned by the previous method. The Damerau–Levenshtein distance is a string metric that calculates the distance between two strings using the operations of insertion, deletion, substitution, and the transposition of two adjacent character. Thus, we could match the professions to the most similar code in order to normalize them.

5 Results

The results were measured using the metrics of Precision, Recall and F1 in Micro-Average. In the experiments, the BETO classifiers were trained for 8 epochs with 80% of the training set for each class independently, and evaluated over 20% of the training set using early stopping criteria.

Table 2 shows the results of the EdIE-KnowLab team for the Task 1 including the performance by entity type. Similarly to the final results, the BETO model using all the training data (BETO+All) obtained good results for the *Profesion* class, but it could not recognize the *Situacion Laboral* and *Actividad* classes. Introducing the undersampling technique which removed the documents without entities (BETO+Positive) shows an improvement in all classes and in the Micro-Average performance. Adding the ProfNER training and validation sets to the BETO model with all the training (BETO+All+ProfNER) increases the performance in *Profesion* by 1.1% in F1, but it still has 0% for the remaining entity types. Surprisingly, the results with the BETO model using only the positive examples of the training set and the ProfNER sets (BETO+Positive+ProfNER) do not overcome the performance of the model trained without ProfNER documents.

Table 2. EdIE-KnowLab team final submissions and their results for Task 1. Best performance for each measure is marked in bold.

Submission		*Profesion*	*Situacion Laboral*	*Actividad*	**Micro-Average**
BETO+All	*Precision*	68.4%	0%	0%	68.4%
	Recall	67.1%	0%	0%	43.3%
	F1	67.7%	0%	0%	53%
BETO+Positive	*Precision*	70%	**45.3%**	**18.6%**	58.6%
	Recall	**76%**	**65.5%**	28.6%	**71.3%**
	F1	**72.8%**	**53.5%**	22.5%	**64.3%**
BETO+All +ProfNER	*Precision*	**71.8%**	0%	0%	**71.8%**
	Recall	66%	0%	0%	42.5%
	F1	68.8%	0%	0%	53.4%
BETO+Positive +ProfNER	*Precision*	62%	42.8%	18.2%	53.5%
	Recall	71.8%	56%	**35.7%**	65.7%
	F1	66.5%	48.5%	**24.1%**	58.9%

Table 3 shows the results of the EdIE-KnowLab team for the Task 2 including the performance by entity type. Similarly to Task 1, the performance in the type *Familiar* is 0% in F1 if the complete training set is used (BETO+ALL) and this could be increased if we only used the positive examples (BETO+Positives) to 16.6% in F1. However, the results in the types *Paciente*, *Sanitario*, and *Otros* are lower resulting in a worse performing Micro-Average. It can be concluded that the models which use the complete dataset have better Precision while the models which only uses the positive examples have better Recall in Micro-Average and almost all the classes.

Table 4 shows the results of the EdIE-KnowLab team for the Task 3 which is the normalization of the recognized mention. Contrary to the previous tasks, the normalization task depends directly on the performance of the recognition

Table 3. EdIE-KnowLab team final submissions and their results for Task 2. Best performance for each measure is marked in bold.

Submission		*Paciente*	*Sanitario*	*Familiar*	*Otros*	**Micro-Average**
BETO+All	*Precision*	**52.9%**	**75.4%**	0%	**64.6%**	**60.4%**
	Recall	**59.3%**	78.2%	0%	50%	60.3%
	F1	**55.9%**	**76.8%**	0%	**56.4%**	**60.4%**
BETO+Positive	*Precision*	50.6%	68.6%	**11%**	28.8%	45.5%
	Recall	54.4%	**84%**	**34%**	**70.5%**	**63.6%**
	F1	52.4%	75.5%	**16.6%**	40.9%	53%

of mentions by the NER models. Thus, the models that obtained better results in the development sets were chosen for normalization of their outputs. For this reason, we decided to use the outputs of the BETO+Positive and the BETO+Positive+ProfNER from Task 1 and the BETO+Positive from Task 2. The results confirms that the performance in the normalization is directly related with the performance in NER tasks. The BETO+Positive from Task 1 is the best configuration reaching 17.8% in Micro-Average F1.

Table 4. EdIE-KnowLab team final submissions and their results for Task 3. Best performance for each measure is marked in bold.

Submission		**Micro-Average**
BETO *NER*+Positive	*Precision*	**16.5%**
	Recall	**19.3%**
	F1	**17.8%**
BETO *NER*+Positive+ProfNER	*Precision*	15.1%
	Recall	17.7%
	F1	16.3%
BETO *CLASS*+Positive	*Precision*	13.4%
	Recall	16.5%
	F1	14.8%

6 Conclusions

This paper presents the participation of the EdIE-KnowLab for the MEDDOPROF recognition shared tasks from IberLeF 2021. The proposed methods are a single BETO classifier applied to each entity type for the classification of occupation mentions and the recognition of the person who the occupation is referring, and the use of the Damerau-Levenshtein distance to match the recognized professions to one of the valid codes from a given list. The NER model obtained

64.3% using only the positives examples, and 60.4% using all the training set in Micro-Average F1 for the Task 1 and 2, respectively. As BERT models usually do not allow sentences greater than 512 tokens, the documents were divided into different parts and the recognized mentions in each split were merged in the final prediction file. In general, the models that used the complete dataset for training obtained best Precision while the models that used only the positive examples obtained best Recall. In the normalization task, the approximate string matching approach reached to 17.8% in F1 with the BETO+Positives outputs of the Task 1. As future work, we will explore the combination of the BETO classifier with different configurations for the NER tasks, and include some preprocessing over the recognized entities such as removing stop words, lemmatization, or stemming for the normalization task.

Acknowledgments

The authors would like to thank to members in the Clinical Natural Language Processing Research Group and KnowLab in the University of Edinburgh and University College London for their valuable discussion and comments. This work was supported by the HDR UK National Text Analytics Implementation Project, Wellcome Institutional Translation Partnership Awards (PIII029), a Legal and General PLC (research grant to establish the independent Advanced Care Research Centre at University of Edinburgh). Legal and General PLC had no role in conduct of the study, interpretation or the decision to submit for publication. The views expressed are those of the authors and not necessarily those of Legal and General PLC.

References

1. Ahmed, A., Abbasi, A., Eickhoff, C.: Benchmarking modern named entity recognition techniques for free-text health record de-identification **abs/2103.13546** (2021), https://arxiv.org/abs/2103.13546
2. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings. AMIA Symposium pp. 17–21 (2001), https://pubmed.ncbi.nlm.nih.gov/11825149, publisher: American Medical Informatics Association
3. Blanquet, M., Labbe-Lobertreau, E., Sass, C., Berger, D., Gerbaud, L.: Occupational status as a determinant of mental health inequities in French young people: is fairness needed? Results of a cross-sectional multicentre observational survey. International Journal for Equity in Health **16**(1), 142 (aug 2017). https://doi.org/10.1186/s12939-017-0634-7, https://doi.org/10.1186/s12939-017-0634-7
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017). https://doi.org/10.1162/tacl$_{a}$0051, https://www.aclweb.org/anthology/Q17-1010

5. Carreto Fidalgo, D., Vila-Suero, D., Aranda Montes, F., Talavera Cepeda, I.: System description for ProfNER - SMMH: Optimized finetuning of a pretrained transformer and word vectors. In: Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task. pp. 69–73. Association for Computational Linguistics, Mexico City, Mexico (Jun 2021). https://doi.org/10.18653/v1/2021.smm4h-1.11, https://www.aclweb.org/anthology/2021.smm4h-1.11
6. Catelli, R., Casola, V., De Pietro, G., Fujita, H., Esposito, M.: Combining contextualized word representation and sub-document level analysis through bi-lstm+crf architecture for clinical de-identification. Knowledge-Based Systems **213**, 106649 (2021). https://doi.org/https://doi.org/10.1016/j.knosys.2020.106649, https://www.sciencedirect.com/science/article/pii/S0950705120307784
7. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pretrained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
8. Chilman, N., Song, X., Roberts, A., Tolani, E., Stewart, R., Chui, Z., Birnie, K., Harber-Aschan, L., Gazard, B., Chandran, D., Sanyal, J., Hatch, S., Kolliakou, A., Das-Munshi, J.: Text mining occupations from the mental health electronic health record: a natural language processing approach using records from the clinical record interactive search (cris) platform in south london, uk. BMJ Open **11**(3) (2021). https://doi.org/10.1136/bmjopen-2020-042274, https://bmjopen.bmj.com/content/11/3/e042274
9. Damerau, F.J.: A technique for computer detection and correction of spelling errors. Commun. ACM **7**(3), 171–176 (Mar 1964). https://doi.org/10.1145/363958.363994, https://doi.org/10.1145/363958.363994
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423
11. Eberts, M., Ulges, A.: Span-based joint entity and relation extraction with transformer pre-training. In: ECAI (2020)
12. Fabregat, H., Duque, A., Martínez-Romo, J., Araujo, L.: De-identification through named entity recognition for medical document anonymization. In: Cumbreras, M.Á.G., Gonzalo, J., Cámara, E.M., Martínez-Unanue, R., Rosso, P., Carrillo-de-Albornoz, J., Montalvo, S., Chiruzzo, L., Collovini, S., Gutiérrez, Y., Zafra, S.M.J., Krallinger, M., Montes-y-Gómez, M., Ortega-Bueno, R., Rosá, A. (eds.) Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019. CEUR Workshop Proceedings, vol. 2421, pp. 663–670. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2421/MEDDOCAN_paper_4.pdf
13. Farré-Maduell, E., Lima-López, S., Miranda-Escalada, A., Brivá-Iglesias, V., Krallinger, M.: MEDDOPROF: Codes reference list (2021), http://doi.org/10.5281/zenodo.4722741
14. Farré-Maduell, E., Lima-López, S., Miranda-Escalada, A., Brivá-Iglesias, V., Krallinger, M.: MEDDOPROF corpus: test set [data set] (2021), http://doi.org/10.5281/zenodo.4889777

15. Farré-Maduell, E., Lima-López, S., Miranda-Escalada, A., Brivá-Iglesias, V., Krallinger, M.: MEDDOPROF corpus: training set + complementary entities [data set] (2021), http://doi.org/10.5281/zenodo.4775741
16. Farré-Maduell, E., Lima-López, S., Miranda-Escalada, A., Brivá-Iglesias, V., Krallinger, M.: MEDDOPROF guidelines (2021), http://doi.org/10.5281/zenodo.4720833
17. Friedman, C., Hripcsak, G., DuMouchel, W., Johnson, S.B., Clayton, P.: Natural language processing in an operational clinical information system. Nat. Lang. Eng. **1**, 83–108 (1995)
18. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing (2021)
19. Heinzerling, B., Strube, M.: BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://www.aclweb.org/anthology/L18-1473
20. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020). https://doi.org/10.5281/zenodo.1212303
21. Lange, L., Adel, H., Strötgen, J.: NLNDE: the neither-language-nor-domain-experts' way of spanish medical document de-identification. CoRR **abs/2007.01030** (2020), https://arxiv.org/abs/2007.01030
22. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (09 2019). https://doi.org/10.1093/bioinformatics/btz682
23. Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced NLP tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 465–476. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.45
24. Lima-López, S., Farré-Maduell, E., Miranda-Escalada, A., Brivá-Iglesias, V., Krallinger, M.: Nlp applied to occupational health: MEDDOPROF shared task at IberLEF 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. Procesamiento del Lenguaje Natural **67** (2021)
25. Liu, Z., Tang, B., Wang, X., Chen, Q.: De-identification of clinical notes via recurrent neural network and conditional random field. Journal of Biomedical Informatics **75**, S34–S42 (2017). https://doi.org/https://doi.org/10.1016/j.jbi.2017.05.023, https://www.sciencedirect.com/science/article/pii/S1532046417301223, supplement: A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry
26. Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodriguez, H., Martin, J.L., Villegas, M., Krallinger, M.: Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In: IberLEF@SEPLN (2019)
27. Mesa Murgado, A., Parras Portillo, A., López Úbeda, P., Martin, M., Ureña-López, A.: Identifying professions & occupations in health-related social media using natural language processing. In: Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task. pp. 141–145. Association for Computational Linguistics, Mexico City, Mexico (Jun

2021). https://doi.org/10.18653/v1/2021.smm4h-1.31, https://www.aclweb.org/anthology/2021.smm4h-1.31
28. Miranda-Escalada, A., Farré-Maduell, E., Lima-López, S., Gascó, L., Brivá-Iglesias, V., Agüero-Torales, M., Krallinger, M.: The ProfNER shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In: Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task. pp. 13–20. Association for Computational Linguistics, Mexico City, Mexico (jun 2021). https://doi.org/10.18653/v1/2021.smm4h-1.3, https://www.aclweb.org/anthology/2021.smm4h-1.3
29. Qi, Y., Liang, T., Ye, H.: Occupational status, working conditions, and health: evidence from the 2012 China Labor Force Dynamics Survey. The Journal of Chinese Sociology **7**(1), 14 (aug 2020). https://doi.org/10.1186/s40711-020-00128-5, https://doi.org/10.1186/s40711-020-00128-5
30. Saluja, B., Kumar, G., Sedoc, J., Callison-Burch, C.: Anonymization of sensitive information in medical health records. In: IberLEF@SEPLN (2019)
31. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA **17**(5), 507–513 (2010). https://doi.org/10.1136/jamia.2009.001560, https://pubmed.ncbi.nlm.nih.gov/20819853, publisher: BMJ Group
32. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: brat: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations Session at EACL 2012. Association for Computational Linguistics, Avignon, France (April 2012)
33. Suárez-Paniagua, V.: Vsp at meddocan 2019 de-identification of medical documents in spanish with recurrent neural networks. In: IberLEF@SEPLN (2019)
34. Turian, J., Ratinov, L., Bengio, Y.: Word representations: A simple and general method for semi-supervised learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 384–394. ACL '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1858681.1858721
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
36. Volkers, A.C., Westert, G.P., Schellevis, F.G.: Health disparities by occupation, modified by education: a cross-sectional population study. BMC Public Health **7**(1), 196 (Aug 2007). https://doi.org/10.1186/1471-2458-7-196, https://doi.org/10.1186/1471-2458-7-196
37. Yaseen, U., Langer, S.: Neural text classification and stacked heterogeneous embeddings for named entity recognition in SMM4H 2021. In: Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task. pp. 83–87. Association for Computational Linguistics, Mexico City, Mexico (Jun 2021). https://doi.org/10.18653/v1/2021.smm4h-1.14, https://www.aclweb.org/anthology/2021.smm4h-1.14