

SICAMESP: Systems for Interpretable Automatic Classification of Medical Reports in Spanish

Andres Duque^{1,2,*}, Lourdes Araujo^{1,2}, Juan Martinez-Romo^{1,2},
Alejandro Sánchez-de-Castro-Fernández¹, Jorge Fernandez-Hernandez¹ and
Alicia Ramirez-Arrabe¹

¹*Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED). ETS Ingeniería Informática. Madrid, 28040, Spain*

²*Instituto Mixto de Investigación - Escuela Nacional de Sanidad (IMIENS). Madrid, 28029, Spain*

Abstract

Project SICAMESP is devoted to the development of interpretable Natural Language Processing (NLP) methods for the automatic assignment of ICD-10 codes to medical reports written in Spanish. The project addresses a clinically relevant yet computationally challenging scenario: extreme multi-label classification over highly imbalanced data, where decisions must be traceable by health professionals. SICAMESP combines deep learning techniques and architectures, keyphrase-based modeling, evidence extraction, semantic similarity methods, and hybrid decision strategies to obtain accurate and explainable predictions. The project is supported by anonymized hospital reports from several specialties and by public benchmarks such as CodiEsp. Current outcomes include improved keyphrase-based resources and an explainable coding pipeline that has surpassed state-of-the-art models in different benchmarks. SICAMESP aims to strengthen Spanish clinical NLP and support safer, more efficient management of medical information.

Keywords

Clinical text coding, explainable AI, ICD-10, Spanish medical reports

1. Introduction and Motivation

Hospitals generate large volumes of free-text documentation, including discharge summaries, diagnostic reports, procedure descriptions, and follow-up notes. Despite its importance for patient care, administrative management, and clinical research, much of the information contained in these documents remains difficult to exploit computationally because it is expressed in unstructured language. The assignment of ICD-10 codes to these types of medical documents is a key step in transforming such textual content into structured information that can be queried, aggregated, reused, and audited. ICD-10 is the tenth revision of the International Classification of Diseases (ICD) [1], which is the most widely used coding system in hospitals to describe diagnoses and medical procedures.

In practice, however, manual coding is costly, time-consuming, and error-prone. The Spanish ICD-10 ecosystem comprises a very large label space, covering diagnoses and procedures with fine-grained distinctions concerning etiology, severity, anatomical location, and intervention type. As a consequence, automatic coding must deal with an extreme multi-label scenario [2], strong label imbalance, scarce observations for many codes, and substantial lexical variation across clinical specialties. For Spanish, the

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

*Corresponding author.

✉ aduque@lsi.uned.es (A. Duque); lurdes@lsi.uned.es (L. Araujo); juaner@lsi.uned.es (J. Martinez-Romo);
asanchez@lsi.uned.es (A. Sánchez-de-Castro-Fernández); jfernandez@lsi.uned.es (J. Fernandez-Hernandez);
aramirez@lsi.uned.es (A. Ramirez-Arrabe)

🌐 <https://www.uned.es/universidad/docentes/informatica/andres-duque-fernandez.html> (A. Duque);

<https://www.uned.es/universidad/docentes/informatica/m-lourdes-araujo-serna.html> (L. Araujo);

<https://www.uned.es/universidad/docentes/informatica/juan-martinez-romo.html> (J. Martinez-Romo)

🆔 0000-0002-0619-8615 (A. Duque); 0000-0002-7657-4794 (L. Araujo); 0000-0002-6905-7051 (J. Martinez-Romo);

0000-0003-2472-4129 (A. Sánchez-de-Castro-Fernández); 0009-0001-8996-8303 (J. Fernandez-Hernandez);

0009-0001-8550-6275 (A. Ramirez-Arrabe)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

problem is particularly relevant because available resources are fewer than for English, while hospital workflows still depend on accurate coding for reimbursement, epidemiological studies, and quality control [3, 4].

The SICAMESP project (<https://nlp.uned.es/SICAMESP/>) was conceived to address this gap from a perspective that combines performance with interpretability. In medical applications, a highly accurate black-box model is not sufficient by itself. In this context, it is very important to understand why a system recommends a given code, which textual evidence supports the decision, and under which circumstances a prediction should be reviewed or discarded [5]. For this reason, the project focuses on interpretable automatic classification of medical reports in Spanish, placing traceable evidence at the center of the modeling strategy.

2. Background and Scientific Challenges

Recent years have witnessed a significant increase in automatic clinical coding research. Most high-performing systems are based on supervised neural architectures, especially transformer encoders or related deep learning models trained over large clinical collections. These approaches can learn powerful representations of medical language and often provide strong performance when frequent labels are sufficiently represented. However, their effectiveness is constrained by the characteristics of the task itself. Automatic ICD-10 coding is not a standard multi-class problem but an extreme multi-label setting in which each document may receive several labels, many labels are very rare, and clinically meaningful distinctions may depend on a few short spans in long reports [2].

A second challenge lies in the mismatch between realistic deployment conditions and purely supervised learning assumptions. In hospital environments, distributions shift over time, specialties differ in style and terminology, and some codes appear so infrequently that they are absent from the training split or only weakly represented. This issue is visible in the CodiEsp benchmark, where test partitions include codes not seen during training [3]. As a result, a model that only memorizes observed label patterns may yield deceptively optimistic results on frequent labels while remaining unreliable for rare, new, or semantically adjacent codes.

The coding of diagnoses and procedures introduces additional complexity. Diagnosis labels follow a hierarchical logic, whereas procedure labels are multi-axial and may require fine distinctions concerning location, technique, device, or approach. Many code descriptions are compositional, and some of the evidence needed to assign a code is discontinuous, abbreviated, or spread across the report. In clinical narratives, relevant clues may appear as isolated phrases, short noun groups, laboratory references, temporally grounded events, or domain-specific abbreviations. Consequently, treating the report as an undifferentiated text block can obscure the evidence that actually determines the label.

Interpretability is therefore not merely a desirable extra, but a scientific requirement for the project. Attention heatmaps or confidence values alone are usually insufficient to support medical decision making. A more explicit link between recommended codes and concrete report content would be therefore required, ideally through localized evidence spans, statistically meaningful phrases, or semantically grounded matches against code descriptions. This is the perspective adopted in SICAMESP: explanations should be built into the modeling pipeline rather than added after the fact.

The Spanish setting also raises a broader resource question. Compared with English, there are fewer publicly available corpora, fewer benchmark results, and fewer domain-specific language models directly adapted to Spanish clinical language. At the same time, Spanish hospital data offer rich opportunities for methodological work because lexical variation, domain abbreviations, and specialty-specific reporting styles create realistic stress tests for NLP systems. Public campaigns such as CodiEsp [3] have helped standardize evaluation, while previous Spanish studies have shown the promise of transformer-based methods, ensemble approaches, and interpretable keyphrase-driven systems [4, 6, 7].

Finally, the emergence of large language models has reopened the debate about how far general-purpose generative systems can be trusted in clinically sensitive workflows. Although such models may offer better coverage, prompting flexibility, and multilingual transfer, they can also blur the provenance

of their decisions and make error analysis harder. SICAMESP uses this moment as an opportunity: instead of replacing interpretable components with fully generative solutions, the project studies how both paradigms can be combined.

3. Project Setting: Funding, Team and Resources

SICAMESP is funded by the Universidad Nacional de Educación a Distancia (UNED) through the Talento Joven 2023 call, under project 2023-VICE-0029. The project is carried out within the NLP&IR Group of the Department of Languages and Computer Systems at UNED. The current research team includes Andres Duque as principal investigator, together with Lourdes Araujo, Juan Martinez-Romo, Alejandro Sanchez-de-Castro, Jorge Fernandez-Hernandez and Alicia Ramirez-Arrabe. In addition to the university environment, the project benefits from collaboration with the Hospital Universitario Fundación Alcorcón (HUFA), which provides access to clinically relevant use cases and anonymized data.

The empirical basis of SICAMESP is strong from the outset. The project builds on a collection of more than 13,000 anonymized hospital reports already annotated with ICD-10 codes and spanning several medical specialties. This institutional dataset is complemented by public benchmarks, especially the CodiEsp corpus [3], which enables controlled comparison with previous systems under a shared evaluation protocol. Together, these resources allow the project to balance realism and comparability: the hospital dataset supports development under realistic distributional conditions, while CodiEsp provides a public reference scenario for reproducible experimentation.

The project also includes a clear ethical and legal framework. The collaboration with HUFA was approved by the corresponding bioethics procedures, and the research relies on anonymized reports with no processing of personal identifiers. This aspect is essential because any practical deployment of clinical NLP technology must be compatible with privacy constraints, restricted access policies, and medically responsible model validation.

From an infrastructural standpoint, the team works with local workstations, dedicated servers, backup systems, and cloud resources when required by large-scale experimentation. This combination is especially important for current neural approaches, which require repeated fine-tuning, embedding generation, and large-batch evaluation over very large label spaces.

4. Objectives and Research Design

The central hypothesis of SICAMESP is that the careful combination of deep learning models with explicitly interpretable NLP components can produce systems that are both accurate and directly usable by health professionals. This hypothesis is operationalized through three main objectives.

First, the project studies machine learning and deep learning methods for automatic ICD-10 coding in Spanish medical reports. This line includes strong transformer-based encoders, multilingual and monolingual biomedical language models, and, more recently, instruction-tuned or generative large language models. The goal is not only to maximize predictive accuracy, but also to identify which families of neural models are more suitable for the clinical coding task under severe class imbalance and limited evidence for rare labels.

Second, SICAMESP develops interpretable and hybrid coding strategies. This objective is the conceptual core of the project. Instead of treating explainability as a post-hoc visualization layer, the project uses inherently interpretable units such as keyphrases, evidence spans localized through Named Entity Recognition (NER) models, and semantic relations between report fragments and ICD-10 descriptions. These components are combined with supervised classifiers and similarity-based models so that the system can justify predictions through explicit textual evidence.

Third, the project aims to implement and evaluate a usable coding tool. Beyond offline metrics, SICAMESP plans the integration of the developed models into a web-based application that can be tested with specialists. This step is necessary to evaluate not only effectiveness but also practical issues

Table 1

Main work packages and current status.

WP	Status	Scope
WP1	Ongoing	Coordination, meetings, and integration of all research lines.
WP2	Advanced	Deep learning and LLM-based ICD-10 coding experiments.
WP3	Advanced	Improvement of keyphrase extraction, NER models and other interpretable coding components.
WP4	Advanced	Design of hybrid pipelines combining neural and interpretable modules.
WP5	Ongoing	Tool development, testing and technical reports on usability.
WP6	Ongoing	Publications, website, competitions, networking, and transfer actions.

such as confidence calibration, coverage, interaction with human coders, and the real usefulness of explanations in day-to-day workflows.

With these objectives in mind, the main pipeline of a proposed system addressing automatic and interpretable ICD-10 coding should ideally present the following components:

1. **Representation learning for coding:** development and evaluation of neural models able to rank or assign ICD-10 labels from entire reports or clinically salient textual segments.
2. **Evidence-centered interpretation:** extraction of phrases, spans, or entities that support a candidate code, allowing the system to display the textual basis of the recommendation.
3. **Semantic grounding of labels:** use of ICD-10 descriptions, keyphrases, and similarity-based retrieval to improve robustness for underrepresented or unseen labels.

In operational terms, Table 1 summarizes the work packages of the project and their current status.

The project is organized into six main work packages that reflect both its methodological progression and its applied orientation. WP1 covers project coordination and scientific management, including the initial planning and the periodic meetings required to monitor progress across all research lines; this task is ongoing, as expected for the full duration of the project. WP2 is devoted to the study and development of machine learning and deep learning approaches for automatic ICD-10 coding, including recent experiments with large language models; this package is currently at an advanced stage, with several promising systems already developed and tested. WP3 focuses on interpretability-oriented methods, especially the improvement of keyphrase-based approaches, development of NER-based models and related techniques aimed at providing explicit textual evidence for code assignment; this package has also reached an advanced stage and constitutes one of the project’s core methodological contributions. WP4 addresses the design of hybrid systems that combine the predictive strength of neural models with the transparency of interpretable components; this package is also at an advanced stage, and recent results already show the potential of this integration. WP5 concerns the implementation and evaluation of a specialist-oriented tool, including supervised testing by healthcare professionals and the preparation of technical reports on system usability and effectiveness; according to the original schedule, this package is still ongoing rather than completed. Finally, WP6 comprises dissemination and transfer activities, such as publications, participation in scientific events and evaluation campaigns, networking, and the preparation of public-facing resources; this task remains ongoing throughout the project.

Evaluation in SICAMESP is multi-layered. At the predictive level, the project uses standard coding metrics such as Precision, Recall, F1, and MAP, which are common in benchmark campaigns. At the explanation level, the project makes use of evaluation strategies that incorporate information about the textual evidence associated with the selected code, either as a phrase, an entity, or a span. Finally, at the user level, the project plans qualitative and task-oriented analysis with specialists, including the perceived usefulness of recommendations, agreement with explanations, and the effect of the tool on annotation speed and confidence. This design deliberately moves away from purely supervised formulations. In a realistic hospital setting, many labels are rare, some are not observed in the training partition, and some clinical expressions appear in formulations not seen before. Therefore, SICAMESP

Table 2
Selected outputs linked to SICAMESP.

Venue / status	Contribution
SEPLN 2025	Explainable ICD-10 coding pipeline integrating NER, classification, and semantic similarity [8].
ECAI 2024	Multilingual temporal expression normalization for biomedical text, relevant to timeline-aware coding and evidence interpretation[19].
CLEF 2024	Early risk detection experiments using representation-learning techniques applicable to sparse clinical text scenarios. [16]
Under review	Extended coding study covering additional datasets and languages.

emphasizes hybrid pipelines where explicit lexical-semantic evidence complements end-to-end neural prediction.

5. Progress Achieved

Most phases of the project have advanced according to plan, with only minor deviations from the original schedule. The state-of-the-art analysis and the initial exploration of deep models were completed early, which enabled the team to move quickly toward system development. In parallel, the keyphrase-based line inherited from previous work [7] was improved with better extraction procedures and broader empirical testing.

The most representative outcome of SICAMESP is an explainable clinical coding pipeline that unifies named entity recognition, extreme multi-label classification, and semantic similarity enriched with keyphrases [8]. In this system, medical evidence is first localized in the report using Named Entity Recognition models, then mapped to diagnosis or procedure codes by a supervised classifier, and finally complemented with a similarity-based stage able to recover low-confidence or out-of-distribution cases. The resulting architecture improves explainability in two ways: it grounds predictions in localized evidence and it preserves a transparent link between evidence texts, keyphrases, and ICD-10 labels. Experiments on CodiEsp show that this approach outperforms previous systems on standard benchmarks across diagnosis, procedure, and explainability subtasks [6, 9]. This work has been extended to evaluate the proposed pipeline on additional public datasets, including Cantemist [10], FALP [11, 12], and MDACE [13], as well as to assess its applicability to the English language. The promising results obtained are currently under review for publication in a leading journal in the field.

The project has also generated reusable methodological components outside the narrow scope of label assignment. Complementary work has been conducted on rare disease detection [14], mental-health risk identification from text [15, 16, 17, 18], and temporal expression normalization in biomedical documents [19, 20]. These lines are useful because they strengthen the project’s broader expertise in clinical NLP, especially in evidence extraction, representation learning, and medically meaningful text normalization.

Table 2 summarizes a selection of outputs already associated with the project. The goal of this table is not to present an exhaustive list, but to illustrate the balance between core coding research and reusable enabling technology.

6. Dissemination, Expected Impact and Transfer

The dissemination plan of the SICAMESP project includes a project website, publication in JCR-indexed journals, participation in conferences such as ECAI or SEPLN, as well as evaluation campaigns such as IberLEF and CLEF. This strategy is appropriate for a project whose outputs range from algorithms and datasets to software components and experimental methodologies.

The expected scientific impact of SICAMESP is threefold. First, the project contributes with new knowledge on explainable extreme multi-label classification for Spanish clinical text. Second, it consolidates language resources, coding benchmarks, and evidence-centered modeling practices that are still

relatively scarce in Spanish biomedical NLP. Third, it offers a concrete case study on how interpretable and neural methods can be combined rather than opposed. These contributions are especially relevant for other clinical NLP tasks, such as event extraction, phenotyping, and multilingual biomedical normalization, where evidence localization and controlled prediction are equally important.

The expected clinical and organizational impact is also significant. Reliable automatic coding can reduce repetitive workload, support more consistent annotation, facilitate downstream statistics and audits, and improve the reuse of medical information for research. Importantly, the project's emphasis on explanations makes its outputs more compatible with real clinical practice than purely opaque recommendation engines. Even when full automation is not desirable, an interpretable coding assistant can accelerate human work while preserving oversight. This point is central to the philosophy of SICAMESP: the aim is to support coding professionals and clinicians, not to replace them.

Technology transfer is an explicit part of the project. Public dissemination will be combined with the possible protection and exploitation of software generated during development. Social dissemination through university communication channels, radio, and online media is also part of the plan, which reflects the public-health relevance of robust medical information processing.

Next steps include the consolidation of the hybrid coding architecture, the implementation of the specialist-oriented web application, and broader validation across datasets, languages, and coding conditions. These next stages are essential to determine whether the project can move from competitive experimental results to practically useful clinical decision-support technology.

7. Conclusions

SICAMESP addresses an important problem at the intersection of clinical practice and natural language processing: the interpretable automatic assignment of ICD-10 codes to Spanish medical reports. The project is grounded on realistic hospital data, supported by public benchmarks, and oriented toward a hybrid methodology where neural prediction and explicit textual evidence reinforce one another.

The progress achieved so far supports the original hypothesis of the project. Interpretable components such as keyphrases, evidence extraction, and semantic similarity are not merely explanatory add-ons; they can be integrated into competitive coding systems that obtain strong results while remaining auditable. At the same time, experiments with deep and large language models provide the necessary modeling power to tackle highly imbalanced, large-label settings.

As a project communication, this paper highlights SICAMESP as both a research program and an applied clinical NLP effort. Its ultimate contribution is expected to be methodological, practical, and infrastructural: better models for Spanish clinical coding, more trustworthy human-AI interaction, and stronger resources for the broader biomedical NLP community.

Acknowledgments

SICAMESP is funded by the Universidad Nacional de Educación a Distancia (UNED), Spain, through the Talento Joven 2023 call (project 2023-VICE-0029). The project also benefits from collaboration with Hospital Universitario Fundación Alcorcón for the study of anonymized medical reports.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check, paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] O. WHO, International classification of diseases, WHO [Internet] (1992).
- [2] J. Liu, W. Chang, Y. Wu, Y. Yang, Deep learning for extreme multi-label text classification, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017, ACM, 2017, pp. 115–124. URL: <https://doi.org/10.1145/3077136.3080834>.
- [3] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF ehealth 2020, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_263.pdf.
- [4] M. Almagro, R. M. Unanue, V. Fresno, S. Montalvo, Icd-10 coding of spanish electronic discharge summaries: An extreme classification problem, *IEEE Access* 8 (2020) 100073–100083.
- [5] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, *WIREs Data Mining Knowl. Discov.* 9 (2019). URL: <https://doi.org/10.1002/widm.1312>.
- [6] G. López-García, J. M. Jerez, N. Ribelles, E. Alba, F. J. Veredas, Transformers for clinical coding in spanish, *IEEE Access* 9 (2021) 72387–72397. URL: <https://doi.org/10.1109/ACCESS.2021.3080085>.
- [7] A. Duque, H. Fabregat, L. Araujo, J. Martínez-Romo, A keyphrase-based approach for interpretable ICD-10 code classification of spanish medical reports, *Artif. Intell. Medicine* 121 (2021) 102177. URL: <https://doi.org/10.1016/j.artmed.2021.102177>.
- [8] A. Ramirez-Arrabe, A. Duque, J. Martínez-Romo, Unifying named entity recognition and extreme multi-label classification for explainable clinical coding, *Proces. del Leng. Natural* 75 (2025) 41–52. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6737>.
- [9] J. Barros, M. Rojas, J. Dunstan, A. Abeliuk, Divide and conquer: An extreme multi-label classification approach for coding diseases and procedures in spanish, in: Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis, LOUHI@EMNLP 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7, 2022, Association for Computational Linguistics, 2022, pp. 138–147. URL: <https://doi.org/10.18653/v1/2022.louhi-1.16>.
- [10] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results., *IberLEF@ SEPLN* (2020) 303–323.
- [11] F. Villena, P. Báez, S. Peñafiel, M. Rojas, I. Paredes, J. Dunstan, Automatic support system for tumor coding in pathology reports in spanish, Available at SSRN 3982259 (2021).
- [12] F. Villena, P. Báez, S. Peñafiel, M. Rojas, I. Paredes, J. Dunstan, Neoplasm topography and morphology corpus, 2021. doi:10.5281/zenodo.5555432, data set.
- [13] H. Cheng, R. Jafari, A. Russell, R. Klopfer, E. Lu, B. Striner, M. Gormley, Mdace: Mimic documents annotated with code evidence, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2023) 7534–7550.
- [14] A. Duque, L. Araujo, J. Martínez-Romo, M. D. Esteban-Vasallo, M.-F. Domínguez-Berjón, D. Malillos Perez, An integrated approach for rare disease detection and classification in spanish pediatric medical reports, *Scientific Reports* 15 (2025) 37973.
- [15] J. Martínez-Romo, L. Araujo, B. Reneses, Guardian-bert: Early detection of self-injury and suicidal signs with language technologies in electronic health reports, *Comput. Biol. Medicine* 186 (2025) 109701. URL: <https://doi.org/10.1016/j.compbimed.2025.109701>.
- [16] H. Fabregat, D. Deniz, A. Duque, L. Araujo, J. Martínez-Romo, NLP-UNED at erisk 2024: Approximate nearest neighbors with encoding refinement for early detecting signs of anorexia, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 813–824. URL: <https://ceur-ws.org/Vol-3740/paper-76.pdf>.
- [17] J. Fernandez-Hernandez, H. Fabregat, A. Duque, L. Araujo, J. Martínez-Romo, UNED-GELP at

- mentalriskes 2024: Transformer-based encoders and similarity techniques for early risk prediction of mental disorders, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), Valladolid, Spain, September 24, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3756/MentalRiskES2024_paper2.pdf.
- [18] J. Martínez-Romo, J. F. H. Barril, L. Araujo, E. A. de la Cal Marín, UNED-UNIOVI at emospeech-iberlef2024: Emotion identification in spanish by combining multimodal textual analysis and machine learning methods, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), Valladolid, Spain, September 24, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3756/EmoSPEECH2024_paper10.pdf.
- [19] A. Sánchez-de-Castro-Fernández, L. Araujo, J. Martínez-Romo, Generative llms for multilingual temporal expression normalization, in: ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024), volume 392 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2024, pp. 3789–3796. URL: <https://doi.org/10.3233/FAIA240940>.
- [20] A. S. d. Castro, L. Araujo, J. Martinez-Romo, A novel methodology for enhancing cross-language and domain adaptability in temporal expression normalization, *Computational Linguistics* 51 (2025) 1303–1335. URL: <https://doi.org/10.1162/COLL.a.12>.