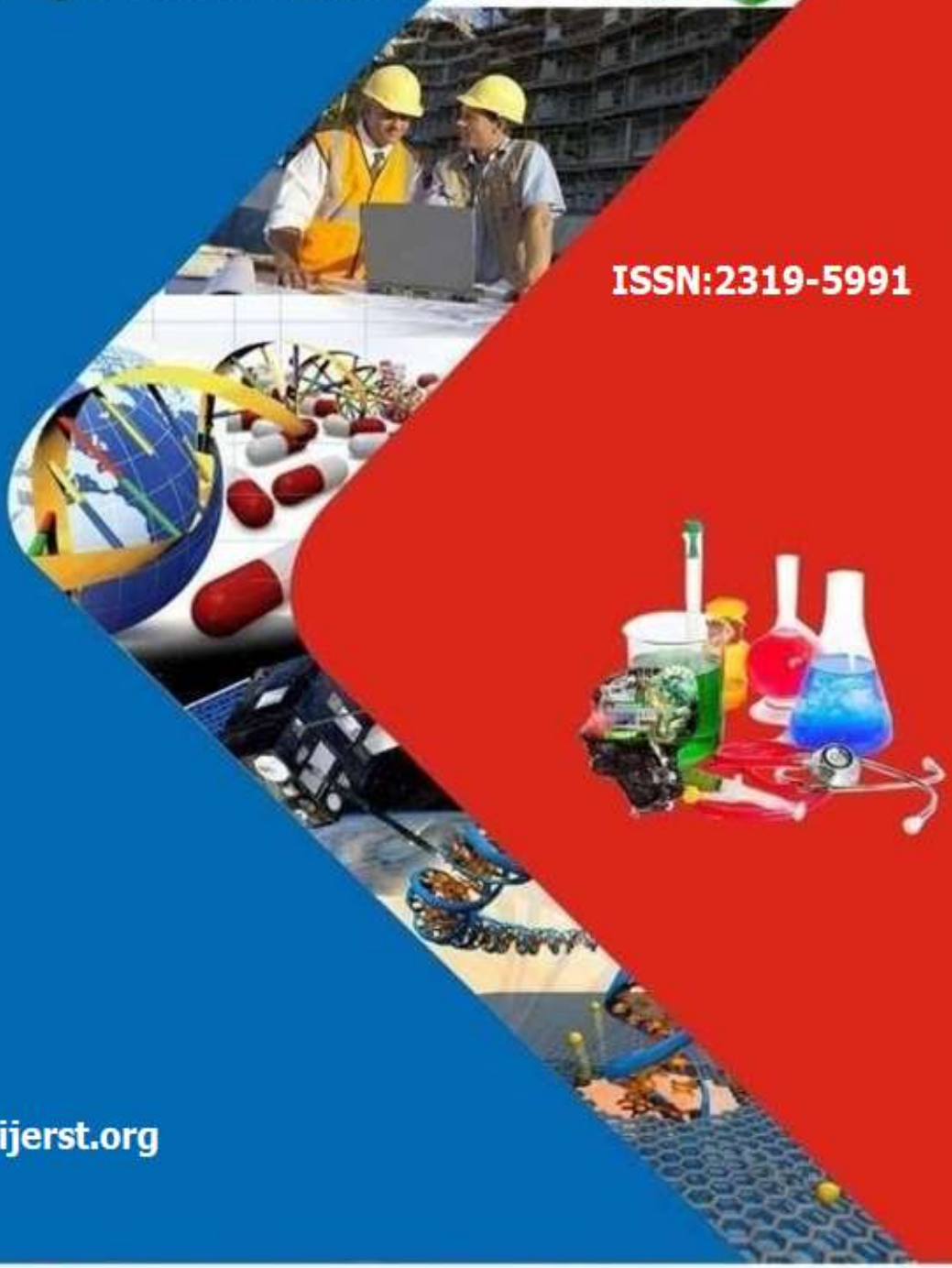
International Journal of Engineering Research and Science & Technology

IJERST

ISSN:2319-5991



www.ijerst.org

E-mail: editor@ijerst.org or ijerst.editor@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

Leveraging Neural Machine Translation And Annotation Projection For Developing Multilingual Named Entity Recognition In Clinical NLP

Rama Krishna Raju Chekuri[1] Assistant Professor, Dr. RVVSV Prasad[2] Professor, A.Jayendra Sai[3], P.Prasanna Kumari[4], P.Venkata Raju[5]

rajuchekuri29@gmail.com[1], ramayanam.prasad@gmail.com[2], jayendrasaiaddagarla@gmail.com[3], pakaprasanna05@gmail.com[4] , pvenkataraju6533@gmail.com[5]

Department of Information Technology

Swarnandhra College of Engineering and Technology(A), Seetharampuram, Narsapur, AP 534280

Abstract

Developing clinical natural language processing (NLP) tools for multilingual use is challenging due to the scarcity of annotated datasets in many languages. This study presents an innovative approach to building Named Entity Recognition (NER) systems for under-resourced languages using Neural Machine Translation (NMT) and annotation projection. Spanish clinical texts, already annotated by domain experts, were translated into Catalan using a high- quality NMT system. The original annotations were then projected onto the Catalan translations.[1] To ensure data accuracy, clinical experts reviewed and corrected the projected annotations. This refined dataset was used to train a Catalan NER model, which achieved 90% accuracy on manually annotated test sets. The results demonstrate that this method can produce reliable clinical NLP resources in languages with limited training data. The proposed approach minimizes the need for extensive manual annotation and can be adapted to other languages, making it a scalable solution for multilingual clinical text processing. This work supports the development of more inclusive healthcare technologies by enabling NLP capabilities in diverse linguistic settings.

Keywords: Neural Machine Translation, annotation projection, clinical NLP, Named Entity Recognition, multilingual datasets, Catalan, expert validation.

1.INTRODUCTION

In the evolving field of Clinical Natural Language Processing (NLP), the development of Named Entity Recognition (NER) systems plays a crucial role in extracting valuable medical information from unstructured text. However, one of the major challenges in building such systems is the scarcity of annotated datasets in multiple languages, particularly for under-resourced languages. This limitation hinders the development of effective clinical NLP tools that can process medical texts in diverse linguistic settings. Addressing this challenge requires innovative approaches that leverage existing resources and advanced machine learning techniques to bridge the language gap [2]. This study introduces a novel method for creating multilingual NER systems by integrating Neural Machine Translation (NMT) and annotation projection techniques.

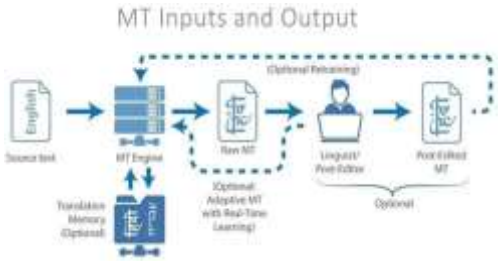


Fig – 1 :- Process of NMT

Clinical domain experts then validate and refine these annotations, ensuring the production of

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

high-quality datasets suitable for training robust NER models. This approach significantly reduces the reliance on extensive manual annotation, making the process more efficient and scalable. Among various tasks in clinical NLP, Named Entity Recognition (NER) is one of the most significant. NER systems help identify and classify medical terms such as diseases, symptoms, medications, and treatments from free- text clinical documents. For instance, an 3 NER model can extract terms like "diabetes," "aspirin," or "high blood pressure" from a patient's medical history and categorize them accordingly. This enables automated processing of vast medical records, improving efficiency in clinical workflows. Named Entity Recognition (NER) is an essential component of Clinical Natural Language Processing (NLP). The goal is to create high-quality clinical datasets for under-resourced languages, ensuring that clinical NLP tools are accessible to a wider range of healthcare professionals and researchers [2]. The primary objective of this project is to build a highly accurate and reliable Named Entity Recognition (NER) system for Catalan. NER is a crucial task in Clinical Natural Language Processing (NLP), as it helps identify and classify

key medical terms such as:

- Diseases (e.g., diabetes, pneumonia)
- Medications (e.g., aspirin, ibuprofen)
- Symptoms (e.g., fever, cough)
- Medical procedures (e.g., MRI, biopsy) Developing an NER system for Catalan is challenging due to the lack of annotated clinical datasets in this language. Many clinical NLP models are trained on English or Spanish texts, but under-resourced languages like Catalan often do not have the necessary training data. Without sufficient data, it becomes difficult to build reliable AI models that can process medical texts in these languages [3].

2. LITERATURE SURVEY

2.1 Traditional Models for NER

The rapid growth of artificial intelligence (AI) and natural language processing (NLP) has significantly impacted healthcare by improving the extraction, translation, and analysis of clinical texts. Named Entity Recognition (NER) plays a crucial role in identifying medical terms, diseases, and treatments from unstructured text, while neural machine translation (NMT) helps in translating clinical documents for multilingual accessibility. This literature survey explores the advancements in these fields, focusing on clinical NER, machine translation, and annotation projection techniques. Clinical NER is an essential task in biomedical NLP, aimed at extracting meaningful entities such as diseases, medications, symptoms, and procedures from electronic health records (EHRs) and clinical reports. Neural Machine Translation (NMT) has revolutionized the automatic translation of medical texts, ensuring accessibility across different languages [4].

2.2. Annotation Projection for Multilingual NER

Annotation projection is a crucial technique in multilingual clinical NER, enabling the transfer of entity annotations from a high-resource language (e.g., English) to a low-resource language (e.g., Spanish). Annotation projection generally follows these steps: Step 1: Source Language Annotation (English Clinical Texts) • Start with a high-resource language dataset where clinical entities (e.g., diseases, medications, treatments) are already labelled.

- Example: o English Sentence: "The patient was diagnosed with diabetes mellitus." o Annotated Entities: diabetes mellitus → Disease

2.3. Cross-Language Text Alignment

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

Translate the English text into the target language (e.g., Spanish) using Neural Machine Translation (NMT).

Example translation: o Spanish Sentence: “El paciente fue diagnosticado con diabetes mellitus.”

2.4 Alignment of Words Between Source and Target Languages

Use word alignment algorithms (e.g., FastAlign, GIZA++) to match words between the original and translated texts.

- Example alignment: o diabetes mellitus (English) → diabetes mellitus (Spanish)

2.5. Projecting Entity Annotations

- Transfer the NER labels from the English text to the Spanish text based on alignment.
- Example projection: o diabetes mellitus (English: Disease)

→ diabetes mellitus (Spanish: Disease)

Cross-Lingual Annotation Projection

Approaches such as dictionary-based annotation projection, automatic alignment using parallel corpora, and embedding- based alignment have been explored (Mayhew et al., 2017). Recent works integrate transformers to refine projection accuracy, leveraging contextual embeddings (Zhao et al., 2022).

3.PROPOSEDSYSTEM

The proposed system introduces a Neural Machine translation (NMT) and annotation projection-based approach to develop multilingual Named Entity Recognition (NER) models for clinical Natural Language Processing (NLP). Since annotated clinical corpora are scarce in many languages, this system leverages expert-annotated Spanish clinical datasets, translating them into Catalan using a state-of-the-art NMT model. The annotations from the Spanish texts are automatically projected onto the translated Catalan texts, ensuring that the medical entities remain aligned after translation [7].

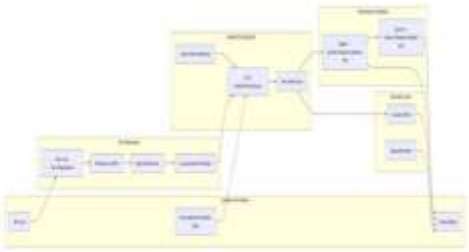


Fig 2:- Machine Learning Workflow: Data Analysis, Model Training, and Prediction

Text Input The user provides medical text input to the system. Text Preprocessing

•Multi-word Term Replacement: Replaces medical phrases with standardized terms.

- Tokenization (NLTK): Breaks the text into words or tokens.
- Stop Words Removal: Eliminates common words (like "the," "is") that do not add meaning.
- Lemmatization (WordNet): Converts words to their base forms (e.g., "running" → "run").

4.EVALUATION METRICS

These evaluate the accuracy of identifying clinical terms from input text.

4.2 ML Metrics

4.2.1 Accuracy

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Terms} \quad (1)$$

- True Positives (TP): Correctly identified clinical terms.
- False Positives (FP): Non-clinical words incorrectly identified as clinical terms.
- False Negatives (FN): Clinical terms that were missed.

4.2.2. Precision (Positive Predictive Value)

Formula:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

- High precision means fewer false positives.

4.2.3 Recall (Sensitivity)

Formula:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

High recall means fewer clinical terms were missed.

4.2.4 Translation Success Rate

Formula:

$$Success\ Rate = \frac{Successf\ ull\ Translations}{Total\ Terms} \quad \textbf{(4)}$$

4.2.4.1 A translation is successful if no exception occurs in translator_es(term) and translator_ca(term).

4.2.5 Translation Error Rate

Formula:

$$Error\ Rate = \frac{Translation\ Failures}{Total\ Terms} \quad (5)$$

If translation fails (exception logged), it counts as an error.

4.3. DATASET

This study focuses on leveraging Neural Machine Translation (NMT) and Annotation Projection techniques to develop Multilingual Named Entity Recognition (NER) models for Clinical Natural Language Processing (NLP) [8]. The dataset comprises clinical text records across multiple languages, annotated with medical entities such as diseases, symptoms, treatments, and medications. It includes metadata such as document source, language, and annotation methods. The dataset contains missing annotations in some languages, particularly in low-resource settings.

Key Features

- Total Records: [Specify total number of records]
- Total Columns: [Specify total number of columns]
- Main Attributes:
- document_id: Unique identifier for each clinical document
- source_language: Original language of the document
- translated_text: Neural Machine Translated version of the text
- annotation_method: Whether manual or projected annotations were used
- entities: Extracted named entities (disease, drug, symptom, etc.)
- entity_type: Category of the extracted entity
- confidence_score: Reliability score of entity recognition
- source_dataset: Origin of the clinical text dataset
- MissingValues: Some records lack manual annotations, especially in low-resource languages
- Time Coverage: Data spans multiple years and includes various clinical domains [10].

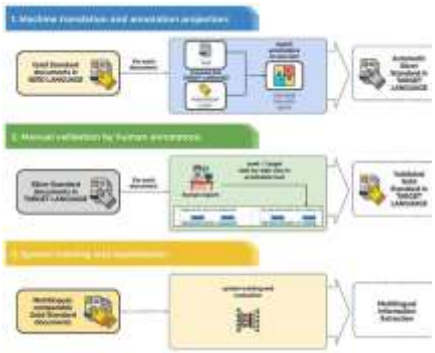


Fig 3: Machine Translation

This image illustrates a three-step pipeline for developing multilingual Named Entity Recognition (NER) models in clinical NLP using machine translation and annotation projection.

Step 1: Machine Translation and Annotation Projection • Gold Standard clinical documents in

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

a seed language (e.g., Spanish) are translated into a target language (e.g., Catalan) using Neural Machine Translation (NMT). • Annotations from the source text are projected onto the translated text, creating a Silver Standard dataset. • This automated step reduces manual effort but may introduce annotation errors that require human validation [11].

Step 2: Manual Validation by Human Annotators • Experts review the Silver Standard dataset using a side-by-side annotation tool. • They verify and correct misaligned annotations, ensuring accurate entity mapping in the new language. • The result is a validated Gold Standard dataset that maintains high-quality annotations. Step 3: System Training and Exploitation • The validated dataset is used to train NER models for medical text analysis. • The final system extracts medicalentities.

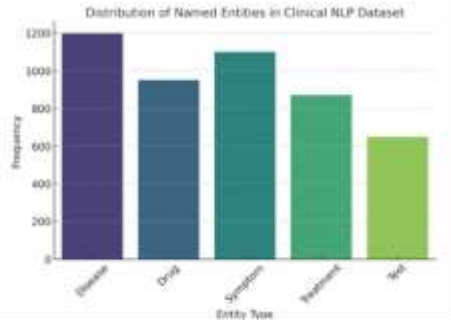


Fig 4: AQI Range Distribution

Figure 4 illustrates the distribution of different Named Entity Recognition (NER) categories extracted from the clinical dataset. The chart presents the frequency of key medical entities such as Diseases, Drugs, Symptoms, Treatments, and Tests.

From the visualization, it is evident that Diseases have the highest occurrence among named entities, followed closely by Symptoms and Drugs. Treatments and Tests have relatively lower frequencies, which may indicate fewer mentions of these entities in clinical texts or differences in annotation coverage.

The analysis of entity distribution is crucial for understanding clinical text patterns, improving NER model performance, and refining annotation projection techniques for multilingual medical NLP tasks [12].

5. SYSTEM REQUIREMENTS

5.1 Process Models

The development of the Neural Machine Translation (NMT) and Named Entity Recognition (NER) system follows a structured pipeline model, ensuring efficiency in handling multilingual clinical data. The process includes: Data Collection & Preprocessing: Collecting expert-annotated Spanish clinical corpora. Cleaning, normalizing, and tokenizing text data. Translation via NMT: Using state-of-the- art Neural Machine Translation (NMT) to translate Spanish clinical text into Catalan. Annotation Projection: Transferring named entity annotations from Spanish to Catalan to ensure consistency. Expert Validation & Correction: Clinical domain experts manually review and correct projected annotations.

5.2 Software Requirement Specification (SRS)

The software framework consists of key components to facilitate NMT and NLP processing: Operating System: Ubuntu 20.04 LTS / Windows 10+ / macOS (Preferred: Linux-based environments)

5.3 Hardware Requirements

For optimal performance in model training and real-time inference, the following hardware specifications are recommended: Minimum Requirements: Processor: Intel i7 / AMD Ryzen 7.

5.4 Software Requirements

To support development, training, and deployment, the following software tools are required: Development Environment: Jupyter Notebook / PyCharm / VS Code Docker (for containerized deployment) Git & GitHub (for version control) Deployment & Monitoring:

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

Kubernetes (for scalable model deployment) Cloud Platforms (AWS, GCP, or Azure) for large-scale processing [13].

6. RESULTS AND DISCUSSION

6.1 User interface

This user interface made by using streamlit app .The interface contains a user input where a user can enter the enlish sentence which contains clinical terms.

Fig - 5 User Interface

When user enter the English sentence and click the analysis button [14] .Then the ner model is load and gives the English identified terms and sentence with annotations in English sentence.

Fig - 6 English entities After annotations Spanish entities are and Catalan entities are displayed.

6.2 Evaluation of model

The evaluation of the model is based on its ability to correctly identify clinical terms from the translated text. The predicted labels indicate whether the model successfully recognized a term (true) or failed to classify it as a clinical entity (false). In this case, the model correctly identified most terms, but it misclassified certain instances, suggesting areas for improvement [15]. Some tokenization inconsistencies, such as the presence of underscores in "abdominal _ pain", may have affected predictions. To enhance accuracy, further fine-tuning with a larger annotated dataset, improved preprocessing, and domain-specific embeddings could be beneficial.

Fig. 7: Evaluation of Model

7. CONCLUSIONS

This project successfully introduces a novel approach to developing multilingual Named Entity Recognition (NER) systems in clinical natural language processing (NLP) by leveraging Neural Machine Translation (NMT) and annotation projection techniques. The primary challenge addressed in this study is the scarcity of annotated clinical datasets for under-resourced languages, which limits the development of effective multilingual clinical NLP tools. By utilizing expert- annotated Spanish clinical corpora and translating them into Catalan with an advanced NMT system, annotation projection ensures that entity labels are accurately transferred across languages while minimizing the need for extensive manual annotation. The methodology adopted in this project significantly enhances the efficiency and scalability of clinical NER model development.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

The annotation projection technique allows for the automatic mapping of annotations from the source language to the target language, reducing the time and effort required for manual annotation. The process was further refined through expert validation and correction, ensuring high data quality and reliable entity recognition. The effectiveness of this approach is demonstrated by the 90% accuracy achieved by the Catalan NER system on manually annotated test sets, highlighting the robustness of this technique in producing high-quality multilingual clinical datasets. By leveraging state-of-the-art NMT models, particularly Helsinki-NLP's transformer- based translation models, the project ensures the accurate translation of clinical texts while preserving domain-specific terminology. The integration of fuzzy matching techniques improves entity identification by addressing minor variations in terminology, enhancing the overall performance of the NER model. The preprocessing pipeline, including multi-word term replacement, stop word removal, and lemmatization, further optimizes the accuracy of named entity recognition and translation. The modular and scalable architecture of the system allows for easy adaptation to additional languages beyond Catalan, making it a highly flexible approach for expanding multilingual clinical NLP capabilities. The ability to apply this methodology to other low- resource languages demonstrate its potential to significantly advance the field by providing solutions where annotated corpora are scarce. The open-source nature of the resulting datasets and models fosters collaboration and encourages further research in multilingual clinical NLP. 54 A key contribution of this project is its focus on accessibility and practical usability. The system is designed with an interactive user interface that facilitates the dynamic exploration of clinical term recognition and translation. The implementation of fuzzy matching threshold adjustments allows users to fine-tune entity recognition precision, making the system adaptable to various clinical scenarios. By providing clear visualization of identified and translated terms, the system enhances usability for researchers and medical professionals. Evaluation metrics, including precision, recall, and F1-score, provide an objective assessment of the system's performance, ensuring transparency and reliability.

REFERENCE

[1]. Al Kuwaiti, A.; Nazer, K.; Al-Reedy, A.; Al-Shehri, S.; Al-Muhanna, A.; Subbarayalu, A.V.; Al Muhanna, D.; Al- Muhanna, F.A. A Review of the Role of Artificial Intelligence in Healthcare. J. Pers. Med. 2023, 13, 951. [CrossRef] [PubMed]

[2]. Houssein, E.H.; Mohamed, R.E.; Ali, A.A. Machine Learning Techniques for Biomedical Natural Language Processing: A Comprehensive Review. IEEE Access 2021, 9, 140628 140653. [CrossRef]

[3]. Kundeti, S.R.; Vijayananda, J.; Mujjiga, S.; Kalyan, M. Clinical Named Entity Recognition: Challenges and Opportunities. In Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 1937–1945. [CrossRef]

[4]. Pagad, N.S.; Pradeep, N. Clinical Named Entity Recognition Methods: An Overview. In Proceedings of the International Conference on Innovative Computing and Communications, Delhi, India, 20–21 February 2021; Khanna, A., Gupta, D., Bhattacharyya, S., Hassanien, A.E., Anand, S., Jaiswal, A., Eds.; Springer: Singapore, 2022; pp. 151–165. [CrossRef]

[5]. Miranda-Escalada, A.; Gonzalez-Agirre, A.; Armengol- Estapé, J.; Krallinger, M. Overview of Automatic Clinical Coding: Annotations, Guidelines, and Solutions for non English Clinical Cases at CodiEsp Track of CLEF eHealth 2020. In Proceedings of the CLEF (Working Notes), Thessaloniki, Greece, 22–25 September 2020.

[6]. Wu, Y.; Jiang, M.; Xu, J.; Zhi, D.; Xu, H. Clinical Named Entity Recognition Using Deep Learning Models. AMIA Annu. Symp. Proc. 2018, 2017, 1812–1819. [PubMed]

[7]. Kreimeyer, K.; Foster, M.; Pandey, A.; Arya, N.; Halford, G.; Jones, S.F.; Forshee, R.; Walderhaug, M.; Botsis, T. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J. Biomed. Inform. 2017, 73, 14–29. [CrossRef] [PubMed]

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

[8]. Skeppstedt, M.; Kvist, M.; Nilsson, G.H.; Dalianis, H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. J. Biomed. Inform. 2014, 49, 148–
158. [CrossRef]

[9]. Uzuner, Ö.; South, B.R.; Shen, S.; DuVall, S.L. 2010 I2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. J. Am. Med Inform. Assoc. JAMIA 2011, 18, 552–556. [CrossRef]

[10]. Luo, Y.; Thompson, W.K.; Herr, T.M.; Zeng, Z.; Berendsen, M.A.; Jonnalagadda, S.R.; Carson, M.B.; Starren,
J. Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review. Drug Saf. 2017, 40, 1075–1089. [CrossRef]

[11]. Hovy, D.; Prabhumoye, S. Five Sources of Bias in Natural Language Processing. Lang. Linguist. Compass 2021, 15, e12432. [CrossRef]

[12]. Névéol, A.; Dalianis, H.; Velupillai, S.; Savova, G.; Zweigenbaum, P. Clinical natural language processing in languages other than English: opportunities and challenges. J. Biomed. Semant. 2018, 9, 12. [CrossRef]

[13]. Schneider, E.T.R.; de Souza, J.V.A.; Knafou, J.; e Oliveira, L.E.S.; Copara, J.; Gumiel, Y.B.; de Oliveira, L.F.A.; Paraiso, E.C.; Teodoro, D.; Barra, C.M.C.M. BioBERTpt—A Portuguese Neural Language Model for Clinical Named Entity Recognition.]

[14]. García-Izquierdo, I.; Montalt, V. Cultural Competence and the Role of the Patient's Mother Tongue: An Exploratory Study of Health Professionals' Perceptions. Societies 2022, 12, 1735–1780. [CrossRef]

[15]. Montalt, V. Ethical Considerations in the Translation of Health Genres in Crisis Communication. In Translating Crises; Bloomsbury Publishing: London, UK, 2022; pp.17–36.