Natural Language Processing (2025), **31**, pp. 56–89
doi:10.1017/nlp.2024.10

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Augmenting a Spanish clinical dataset for transformer-based linking of negations and their out-of-scope references

Antonio Jesús Tamayo-Herrera[1], Diego A. Burgos[2] and Alexander Gelbukh[1]

[1]Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexico and [2]Wake Forest University, Winston-Salem, NC, USA
Corresponding author: Antonio Jesús Tamayo-Herrera; Email: antonio.tamayo@udea.edu.co

(Received 24 June 2023; revised 19 April 2024; accepted 22 April 2024; first published online 17 May 2024)

Abstract

A negated statement consists of three main components: the negation cue, the negation scope, and the negation reference. The negation cue is the indicator of negation, while the negation scope defines the extent of the negation. The negation reference, which may or may not be within the negation scope, is the part of the statement being negated. Although there has been considerable research on the negation cue and scope, little attention has been given to identifying negation references outside the scope, even though they make up almost half of all negations. In this study, our goal is to identify out-of-scope references (OSRs) to restore the meaning of truncated negated statements identified by negation detection systems. To achieve this, we augment the largest available Spanish clinical dataset by adding annotations for OSRs. Additionally, we fine-tune five robust BERT-based models using transfer learning to address negation detection, uncertainty detection, and OSR identification and linking with their respective negation scopes. Our best model achieves state-of-the-art performance in negation detection while also establishing a competitive baseline for OSR identification (Macro F1 = 0.56) and linking (Macro F1 = 0.86). We support these findings with relevant statistics from the newly annotated dataset and an extensive review of existing literature.

Keywords: machine translation; evaluation

1. Introduction

Negation is a linguistic phenomenon that reverses, tones down, or intensifies the truth value of a linguistic unit (proposition, phrase, or word) that undergoes negation (Martí et al., 2016). According to the literature, about half of the natural language statements in the clinical domain feature some sort of negation (Chapman et al., 2001). Negation detection plays a crucial role in clinical text mining by identifying instances where negations modify affirmed clinical entities, including findings, diseases, procedures, body parts, and drugs. When clinical entities, such as symptoms and diagnoses, are negated, their validity is compromised. For instance, phrases like "no cough" or "no fever" indicate the absence of these symptoms (Dalianis, 2018).

A negated statement is made up of three components, namely, the negation cue, the negation scope, and the negation reference, which can be out of the negation scope. While this study focuses on Spanish, Figure 1 shows an example in English for an illustration of this:

Negation cues are understood as words, prefixes, suffixes, or morphemes that indicate negation. A negation scope refers to words that are influenced by a cue. An out-of-scope negation reference is the part of the text to which the negation refers and which is outside its scope.

Table 1. Scenarios of incomplete negated statements

1. Statement without reference:	*. . .did not differ between the two groups.*
2. Statement without cue:	*The incidence of serious adverse events did . . . differ between the two groups.*
3. Statement without scope:	*The incidence of serious adverse events did not . . .*

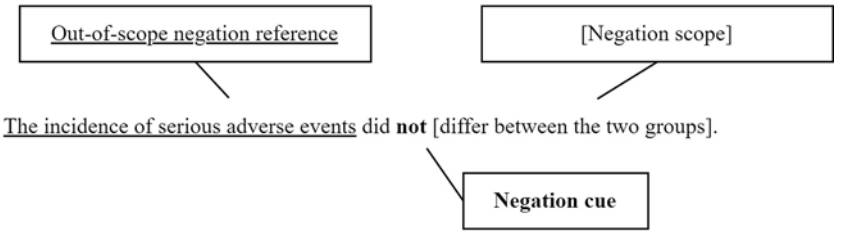

Figure 1. Negated statement components.

The high frequency of negation in electronic health records, clinical trials, drug prescriptions, case studies, discharge summaries, etc. has made researchers investigate automatic methods for its processing and representation for over two decades due to the crucial role of text mining and natural language inference for medical knowledge systems. Negation has gained special attention in the last decade, especially after the emergence of various corpus tagged with negation cues and their corresponding scopes, because of the paramount importance of a correct interpretation of medical information to help reduce medical errors and strengthen decision support. From a computational perspective, the problem has been widely studied in English, and just some years ago a few works emerged to address the issue in Spanish (see, e.g.,, Jiménez-Zafra et al. (2019)), despite being the second language with the most speakers in the world, according to Cervantes Institute. After the works of Vincze et al. (2008) and Morante and Blanco (2012), there appeared new datasets (see a detailed list in Tables 7 and 8 for Spanish) that allow for tackling the problem of identifying negation cues and their scopes automatically, following diverse approaches, which in most cases treat the problem under a token classification scheme.

Negation cues can be easily identified with a list of words or even with a machine learning classifier, but this makes it a domain- or language-dependent task (Fancellu et al., 2016). There are also more complex challenges such as multi-word and discontinuous negation cues, which limit word list performance. Notwithstanding the complexity of the task, negation cue and scope detection have achieved very competitive results. Additionally, the relevance of identifying negations to improve clinical text mining has been shown in many studies (Mutalik, Deshpande, and Nadkarni (2001); Deléger and Grouin, (2012); Santiso et al. (2014); Casillas et al. (2016); Gkotsis et al. (2016). Surprisingly, however, the literature (Morante and Blanco (2021); Mahany et al. (2022)) shows that little to no attention has been paid to identifying negation references located out of the negation scope in spite of the fact that they account for 42.8% of negations, like we show below (see an example of this phenomenon in Figure 1). Considering that a negated statement is a semantic unit, leaving out any of its essential components during its treatment yields senseless, incomplete, or counterfactual chunks. Table 1 illustrates this by using the example in Figure 1 above:

Out of these three scenarios, the only one that actually is an issue in current systems is number 1 in Table 1 above. The practical impact of this flaw on medical data processing is not minor though. Table 2 shows evidence and examples of negation cues and scopes that are currently detected (left column) as well as the information that is not (right column), because it comes in the source text in the form of what we call out-of-scope references (OSRs).

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 2. Examples of truncated negated statements

Negation cue + scope	Truncated information
Not previously known	What was not previously known?
No alterations	What shows no alteration?
No findings	What test showed no findings?
Asymptomatic	Who is asymptomatic?
Do not take Tuesday, Thursday, or Saturday	What is not taken on those days?
Without incident	What was carried out without incident?
Without complications	What was done without complications?
Could not be provided through primary care	What could not be provided?
Without further action	What/who did not keep acting?
Without radiological improvement	What treatment was carried out and what disease did not have a radiological improvement?
Without assistance	Who does what without assistance?
Unable to pinpoint clearly	What is the patient unable to specify?
No pneumonic condensation	Who and in which test did not present pneumonic condensation?
No reports	What is unreported?
No improvement	Who/what had no improvement?

This work's goal is twofold. First, we aim at identifying OSRs to restore the sense of truncated negated statements such as the ones in Table 2 above. For this, we augment the NUBES dataset (Lima et al. 2020) with OSR annotations. NUBES is the largest clinical Spanish dataset available annotated with negations and uncertainty, and now we augment it into NeRUBioS, the first negation and uncertainty clinical Spanish dataset with OSR annotations. We propose that the dataset allows for (*1)* a quantification of the OSR phenomenon and (*2)* using token classification to link negation scopes and their OSRs as most of them are in the same sentence and ambiguity is neglectable (1.2%). Additionally, we fine-tuned five BERT-based models and used transfer learning to jointly identify negation cues and scopes and their respective OSRs as well as uncertainty. Our best model keeps state-of-the-art performance at negation and uncertainty detection and, at the same time, sets a competitive baseline for OSR identification and linking.

The remainder of the paper is as follows. Section 2 presents a survey of related works grouped by the approaches used for negation and uncertainty/speculation detection. This section also includes the details of the most prominent datasets used in the field. Section 3 describes our dataset details as well as our methodology for the dataset annotation process. We provide fine-grained statistics of NeRUBioS. This section also features a description of the transfer learning experiments to jointly tackle negation and uncertainty detection as well as OSR identification and linking. Results and discussion are reported in Section 4. Finally, Section 5 wraps up with conclusions.

2. Related works

According to our review of the field, the present work is the first contribution to tackle the OSR detection and linking problem; hence the lack of related works about this specific task in this

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

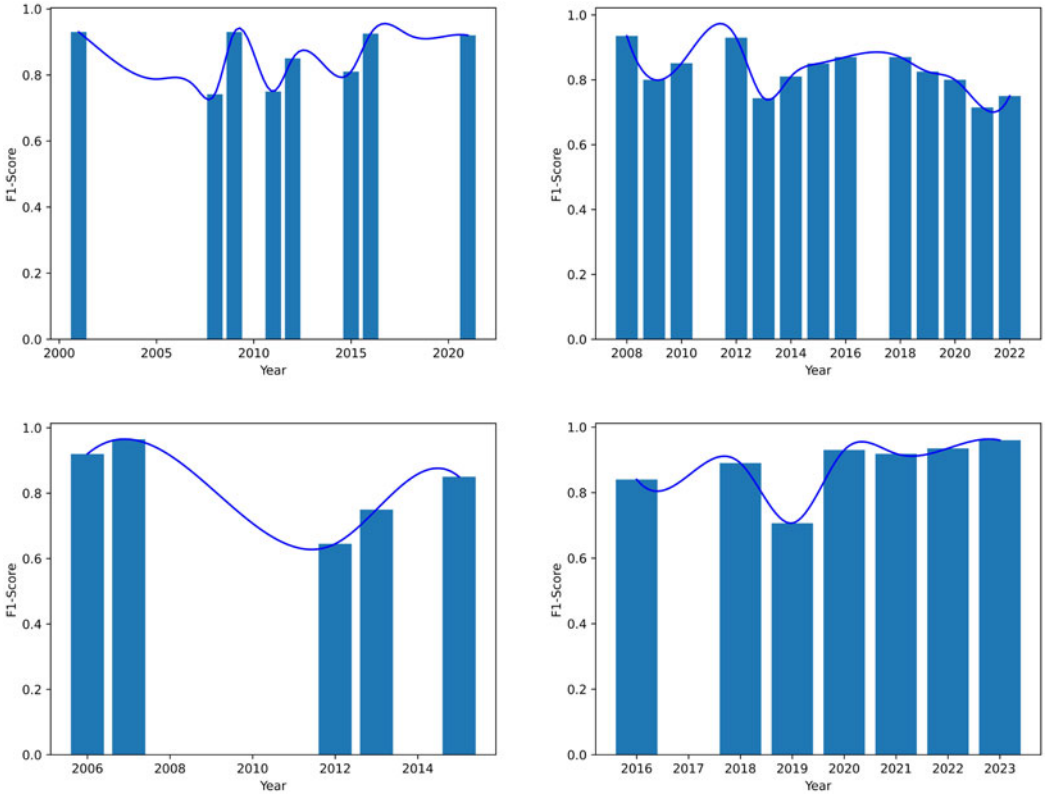

Figure 2. Systems performance by approach over time. Rule-based (upper left), machine learning-based (upper right), hybrid (lower left), and deep learning-based (lower right).

section. However, as OSRs are necessarily tied to negated statements, this study builds on a long tradition of works on negation detection.

Our survey shows that two particular aspects seem to determine the performance of negation detection systems, namely the difficulty of the problem and the language of the dataset. The task has become more complex over time. While the first systems focused mostly on negation cues, later, more sophisticated works added negation scopes and uncertainty cues and scopes to the task. This is probably why researchers achieve high F1 scores at the beginning of a systems' generation, whatever the year or the approach is. For example, a rule-based approach in 2001 (Mutalik et al., 2001) reports an F1 score of 0.96, and 22 years later, Mahany, Khaled, and Ghoniemy (2023) also report an F1 score of 0.96 using an architecture that combines deep-learning techniques.

Systems seem to lose power as awareness of the problem's nuances and challenges increase, though. This can be seen in a consistent decrease in systems' performance by approach over time, except for modern deep-learning systems, which seem to keep high performance. Figure 2 shows this for the works listed in this section's tables by approach.

With these generalities in mind, we briefly discuss each of the tables below, which list relevant works by approach in chronological order.

2.1 Rule-based approaches

By 2000, part-of-speech tagging had reached competitive performance. This enabled ruled-based approaches to use regular expressions, syntactic, and dependency parsing and also lists of negation

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 3. Related works using rule-based approaches

System/Author	Features/Method	Dataset	Performance
NegFinder Mutalik et al. (2001)	Context free grammar parser and Regex	Surgical notes and discharge summaries	$F1 = 0.96$
NegEx Chapman et al. (2001)	Regex	Discharge summaries	$F1 = 0.90$
NegHunter Gindl et al. (2008)	Syntactical approach	Clinical Practice Guidelines	*Rec.* $= 0.83$ *Prec.* $= 0.67$
ConText Harkema et al. (2009)	Negated conditions	Clinical reports	$F1 = 0.93$
ScopeFinder Apostolova et al. (2011)	Lexical and syntactic patterns	BioScope Corpus	**Negation** $F1 = 0.75$ **Speculation** $F1 = 0.76$
Ballesteros et al. (2012)	Syntactic patterns, dependency analysis	BioScope Corpus	$F1 = 0.77$ (Papers) $F1 = 0.85$ (Abstracts) $F1 = 0.93$ (Clinical)
DEEPEN Mehrabi et al. (2015)	Regex, dependency analysis	Clinical notes	$F1 = 0.81$
NegEx Cotik et al. (2016) ∗German	Syntactic patterns, dependency analysis	GNSC	**Negation** $F1 = 0.94$ **Speculation** $F1 = 0.42$
Gkotsis et al. (2016)	Syntactic features	6,066 sentences from Clinical Record Interactive Search (CRIS) dataset	$F1 = 0.91$
Solarte-Pabón et al. (2021b) ∗Spanish	Regex	Nubes dataset	**Negation** *cue* $= 0.95$ *scope* $= 0.89$ **Uncertainty** *cue* $= 0.90$ *scope* $= 0.85$

cues for negation detection. Biomedical datasets were not as big as they are today and they were available mainly in English, with the exception of Cotik et al. (2016) and Solarte-Pabón et al. (2021b). Speculation was introduced by Apostolova, Tomuro, and Demner-Fushman (2011) and Cotik et al. (2016) following this approach (see Table 3). The decreasing trend in these systems' performance over time (see Figure 2) reflects on the increasing complexity of the problem, which probably motivated the next generation of works using machine learning. It is interesting to note that rule-based approaches were still being used for years after the first machine learning works were reported for the task in 2008, as shown below.

2.2 Classic machine learning

The complex configuration of negation and speculation scopes as well as increasing computational capabilities encouraged the use of machine learning techniques (see Table 4). In this generation, speculation has more prominence alongside negation, and more works in Spanish are reported. However, speculation detection is not as promising yet. Most techniques in this approach involve supervised learning using linguistic features. These features are often generated using the data and

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 4. Related works using classic machine learning algorithms

Author/Language	Features/Method	Dataset	Performance
Rokach et al. (2008)	Cascading classifiers, decision trees	Discharge summaries	$F1 = 0.96$
Morante et al. (2008)	K-nearest neighbors and a customized similarity metric	Part of the Bioscope corpus	**Negation cue** $MicroF1 = 0.94$ **Negation scope** $MicroF1 = 0.88$
Morante and Daelemans (2009)	PoS, Lemma, Syntactic SVM, CRF, TiMBL	Bioscope corpus	$F1 = 0.80$
Agarwal and Yu (2010)	CRF	Bioscope corpus	**Negation cue** $F1 = 0.97$ (Papers) $F1 = 0.98$ (Clinical) **Negation scope** $F1 = 0.85$ (Papers) $F1 = 0.95$ (Clinical)
Councill et al. (2010)	Negation cues, dependency analysis, PoS CRF	Product review	$F1 = 0.80$
Zhu et al. (2010)	Shallow semantic parsing	Bioscope corpus SVM	**Negation cue** $F1 = 0.95$ (Abstracts) $F1 = 0.89$ (Papers) $F1 = 0.88$ (Clinical) **Speculation cue** $F1 = 0.88$ (Abstracts) $F1 = 0.77$ (Papers) $F1 = 0.49$ (Clinical) **Negation scope** $F1 = 0.79$ (Abstracts) $F1 = 0.57$ (Papers) $F1 = 0.81$ (Clinical) **Speculation scope** $F1 = 0.77$ (Abstracts) $F1 = 0.50$ (Papers) $F1 = 0.37$ (Clinical)
Cruz Diaz et al. (2012)	Syntactic features Naïve Bayes, Decision tree (C4.5) and SVM	BioScope Corpus	**Negation** $F1 = 0.93$ **Speculation** $F1 = 0.81$
Zou et al. (2013)	Syntactic features SVM Tree kernel	Bioscope corpus	**Negation scope** $F1 = 0.77$ (Abstracts) $F1 = 0.61$ (Papers) $F1 = 0.85$ (Clinical) **Speculation scope** $F1 = 0.84$ (Abstracts) $F1 = 0.67$ (Papers) $F1 = 0.73$ (Clinical)
Wu et al. (2014)	Syntactic features SVM	SHARPn MiPACQ i2b2 corpus NegEx test set	$F1 = 0.98$ $F1 = 0.74$ $F1 = 0.94$ $F1 = 0.58$

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 4. Continued

Author/Language	Features/Method	Dataset	Performance
Attardi et al. (2015)	Dependency analysis SVM, Conditional Markov Model, Neural Networks	Bioscope corpus	**Negation and speculation cue** $F1 = 0.91$ **Negation and speculation scope** $F1 = 0.79$
Cruz et al. (2016)	Dependency analysis SVM + post-processing with syntactic rules	SFU corpus	**Negation cue** $F1 = 0.90$ **Speculation cue** $F1 = 0.92$ **Negation scope** $F1 = 0.84$ **Speculation scope** $F1 = 0.79$
Loharja et al. (2018) Spanish	CRF	SFU ReviewSP NEG	**Negation cue** $F1 = 0.87$
Beltrán and González, (2019) Spanish	CRF	SFU ReviewSP NEG	**Negation cue** $F1 = 0.84$
Domínguez-Mas et al., (2019) Spanish	CRF, Random Forest, SVM, XGBoost	SFU ReviewSP NEG	**Negation cue** $F1 = 0.81$
Jiménez-Zafra et al. (2020a) Spanish	CRF	SFU ReviewSP NEG	**Negation cue** $F1 = 0.87$ **Negation scope** $F1 = 0.73$
Bel-Enguix et al. (2021) Spanish	CRF	T-MexNeg SFU SFU ReviewSP NEG	**Negation scope** $F1 = 0.75$ **Negation scope** $F1 = 0.68$
Tamayo et al. (2022a) Spanish	Syntactic features CRF	SFU ReviewSP NEG	**Negation scope** $MacroF1 = 0.75$

the knowledge built through rule-based techniques during the first years of research on negation detection.

2.3 Hybrid approaches

As classic machine learning seemed to lag behind new challenges in negation detection, hybrid proposals emerged (see Table 5). This approach consists of mixing machine learning and some sort of pre- or post-processing or reinforcement learning. All of the works reported here use English, and, for some reason, they leave out uncertainty/speculation. Overall results were not as big a jump as expected compared to machine learning-only approaches, which led the scientific community to move to the next paradigm, which we describe below.

2.4 Deep learning

Deep learning applied to language processing through large language models (LLM) truly created a new paradigm. While the rule-based, machine learning, and hybrid approaches cited here

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 5. Related works based on hybrid approaches

Author	Features/Method	Dataset	Performance
Goryachev et al. (2006)	Regex + Naïve Bayes or SVM	Discharge reports from Boston-based hospitals	*Accuracy* = 0.92
Huang and Lowe (2007)	Regex + Grammatical análisis	Radiology reports	*Sensitivity* = 0.93 *Specificity* = 1.0
Gyawali and Solorio (2012)	SVM + regex	SEM 2012∗	**Negation cue** $F1 = 0.86$ **Negation scope** $F1 = 0.76$
White (2012)	Regex + CRF	SEM 2012∗	**Negation scope** $F1 = 0.48$
Fujikawa et al. (2013)	Statistics + Heuristics Parse Tree IGTree (NegFinder)	Bioscope corpus	**Negation scope** $F1 = 0.77$ (Abstracts) $F1 = 0.62$ (Papers) $F1 = 0.86$ (Clinical)
Reitan et al. (2015)	CRF + lexicon-based pattern matching	Twitter data	**Negation scope** $F1 = 0.85$
Pröllochs et al. (2017)	Reinforcement learning	IMDB dataset	$R^2 = 0.22$

overlap in time to a certain extent to compete for good results in negation detection, deep learning has drawn almost exclusive interest from the community due to consistent and improving results on this and other tasks over time (Table 6). This can be seen in Figure 3, which shows all of the related works listed above by year. The Figure shows, for example, that there is a publication time overlap of 8 years between rule-based and machine learning and of 6 years between machine learning and hybrid methods. However, there is only a time overlap between classical machine learning and deep-learning publications of 6 years and virtually no overlap with the hybrid generation. This overlap, however, could be even smaller if we consider that the early works that we cite do not exactly use LLMs but word embeddings and other neural network algorithms that came before the advent of BERT and related LLMs.

The number of deep-learning works in Spanish and other languages has increased dramatically, as can be seen in Jiménez-Zafra et al. (2020), who did a thorough review of datasets for negation in Spanish and other languages including this and other approaches. This is mostly due to the availability of models that have been trained on multilingual corpora, such as mBERT (Devlin et al. 2018) (see e.g., Solarte-Pabón et al. (2021a)), or on datasets in the specific language. Table 7 shows a sample of the most representative datasets in Spanish for the clinical/biomedical domain. Some of these datasets are not currently annotated with negation labels, but we include them here for reference as they can potentially be used for further dataset construction. Likewise, Table 8 lists Spanish datasets used for negation detection in other domains. The third column in each table highlights in bold when the dataset contains negation or uncertainty annotations, besides other types of labels, as well as the number of labels per type.

3. Methodology

This work follows a data-based approach which is at the core of transfer learning using large language models. Therefore, in this section we describe the dataset used and its annotation process as well as the five pre-trained models tested and the fine-tuning parameters utilized to optimize the results for each of the tackled tasks.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 6. Related works using deep learning

Author/Language	Features/Method	Dataset	Performance
Fancellu et al. (2016)	Word embeddings and PoS Multilayer perceptron and bidirectional LSTM	SEM 2012∗	$F1 = 0.89$
Qian et al. (2016)	Convolutional neural networks	Bioscope corpus	**Negation scope** F1 = 0.77 (Abstracts) F1 = 0.55 (Papers) F1 = 0.90 (Clinical) **Speculation scope** F1 = 0.86 (Abstracts) F1 = 0.60 (Papers) F1 = 0.74 (Clinical)
Fabregat et al. (2018) Spanish	BiLSTM	SFU ReviewSP NEG	**Negation cue** (computers domain omitted) $F1 = 0.89$
Lazib et al. (2019)	Recurrent neural networks	SFU review	$F1 = 0.89$
Fabregat et al. 2019a) Spanish	Convolutional networks + LSTM	Bioscope corpus	**Negation cue** $F1 = 0.98$ (Abstracts) $F1 = 0.90$ (Papers) $F1 = 0.99$ (Clinical) **Negation scope** $F1 = 0.80$ (Abstracts) $F1 = 0.50$ (Papers) $F1 = 0.95$ (Clinical) **Negation cue** $F1 = 0.96$ **Negation scope** $F1 = 0.72$
Fabregat et al., 2019b) Spanish	BiLSTM + post-processing	SFU ReviewSP NEG	**Negation cue** $F1 = 0.83$
Giudice (2019) Spanish	Convolutional recurrent neural network	SFU ReviewSP NEG	**Negation cue** $F1 = 0.23$
Khandelwal and Sawant (2019)	BERT (NegBERT)	Bioscope corpus SEM 2012∗ SFU Review Corpus	**Negation scope** $F1 = 0.93$ F1 = 0.92 F1 = 0.90
Fei et al. (2020) English, Chinese	Recursive neural networks + CRF	CNeSp	**Negation scope** $F1 = 0.93$ **Speculation scope** $F1 = 0.91$
Lima et al. (2020) Spanish	BiLSTM	Nubes	**Negation** $cue = 0.95$; $scope = 0.91$ **Uncertainty** $cue = 0.85$; $scope = 0.79$
Fabregat et al. (2021) Spanish	Deep neural networks	SpRadIE dataset	**Negation scope** $F1 = 0.95$ **Speculation scope** $F1 = 0.72$

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 6. Continued

Author/Language	Features/Method	Dataset	Performance
García-Lago and Segura-Bedmar (2021) Spanish	BERT BiLSTM + CRF CRF	SpRadIE dataset	**Negation scope** F1 = 0.94 **Speculation scope** F1 = 0.73
Hartmann and Søgaard (2021) Multilingual	Multilingual BERT	A sample of Nubes	**Negation scope** *F-score* = 0.96
López-Úbeda et al. (2021) Spanish	BERT	SpRadIE dataset	**Negation scope** *F*1 = 0.85 **Speculation scope** *F*1 = 0.75 approx.
Polignano et al. (2021) Spanish	Transformers + CRF	SpRadIE dataset	**Negation scope** F1 = 0.90 **Speculation scope** F1 = 0.5 approx.
Solarte-Pabón et al. (2021a) Spanish	Multilingual BERT	SpRadIE dataset	**Negation scope** F1 = 0.89 **Speculation scope** F1 = 0.58
Suárez-Paniagua et al., (2021) Spanish	Ensemble of BERT-like models	SpRadIE dataset	**Negation scope** F1 = 0.94 **Speculation scope** F1 = 0.53
Pabón et al. (2022) Spanish	BiLSTM-CRF Multilingual BERT	Nubes dataset	**Negation** *cue* = 0.95; *scope* = 0.92 **Uncertainty** *cue* = 0.84; *scope* = 0.80
Mahany et al. (2023) Arabic	BERT + BiLSTM + CRF (AraBERT)	ArNegSpec corpus	**Negation cue** F1 = 0.98 **Speculation cue** *F*1 = 0.98 **Negation scope** *F*1 = 0.94 **Speculation scope** *F*1 = 0.95

3.1 The dataset

Fine-tuning a pre-trained language model requires supervised learning, hence the need of a dataset annotated for the relevant task. While there are datasets available for negation and uncertainty detection, there is no dataset for OSR detection and linking. In this section we describe how we augmented an existing dataset with OSR annotations and report the experiments carried out with fine-tuned, multi-task language models for negation detection, uncertainty detection, and OSR identification and linking.

For the OSR identification and linking task, we created NeRUBioS, an augmented version of NUBES (Lima et al., 2020), which is the largest publicly available corpus annotated for negation and uncertainty in Spanish in the biomedical domain. NeRUBioS (Negation, References, and Uncertainty in Biomedical Texts in Spanish), the first dataset of its kind, resulted from manually annotating NUBES with OSR tags (see Figure 4).

NeRUBioS contains 18,437 sentences from anonymized health records, which total 342,788 tokens and 32,562 types. 7,520 of these sentences contain negated statements, out of which 3,606

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

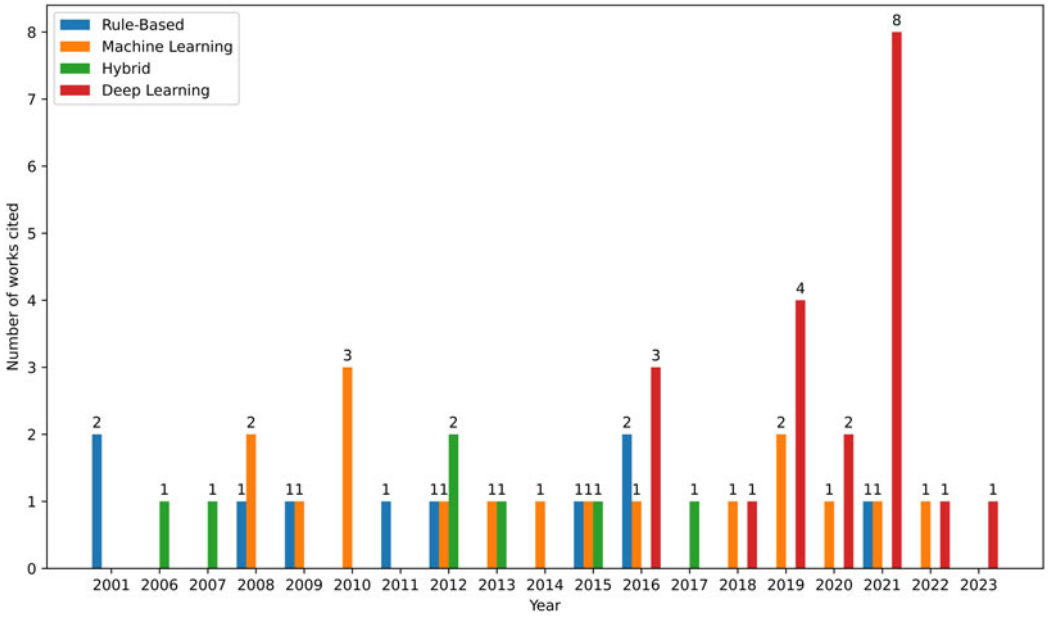

Figure 3. Approach overlaps in publications over time.

(47.9%) feature OSRs. Figure 5 shows the number of sentences per partition. We follow Lima et al. (2020) for this distribution of samples across partitions in order to allow for a comparative analysis of each task using NUBES vs. NeRUBioS, that is, 75% of the dataset for training, 10% for development, and 15% for testing. For more statistics about NeRUBioS, see Table 9 below.

3.2 Tagset

In order to be consistent with the annotation scheme inherited from NUBES, NeRUBios' OSRs were tagged using the BIO scheme (Ramshaw and Marcus, 1999). Most OSRs and their negation scopes are in the same sentence, which enables the training of a model to make the linking inference between a negation and its OSR by tackling the task as a token classification problem.

After annotating every OSR, the dataset ends up having 11 labels (See Table 10 above). The bilingual example below illustrates the use of this tagging (Spanish version in italics):

labrum	*anterior*	*y*	*posterior*	*no*	*presentan*	*alteraciones*	.
B-NegREF	I-NegREF	I-NegREF	I-NegREF	B-NEG	B-NSCO	I-NSCO	O
anterior	and	posterior	labrum	show	no	alterations	.
B-NegREF	I-NegREF	I-NegREF	I-NegREF	B-NSCO	B-NEG	I-NSCO	O

Figures 6 and 7 show the frequency of each label across the three partitions of the dataset. Due to space limitations, the Figures do not show the frequencies of the label "O," which are as follows: training = 202,364, development = 27,927, and testing = 39,674. Likewise, Figures 6 and 7 highlight the imbalance in the number of labels in the dataset, which adds to the complexity of each task. The figures unveil some relevant facts per dataset partition, for example, that most of the OSRs (NegREF) are multi-word sequences while most of the negation cues (NEG) are one-word strings. Negation scopes (NSCO) are way longer than the rest of the classes.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 7. Spanish clinical/biomedical datasets

Dataset	Details	Annotations
MultiMedica Moreno-Sandoval and Campillos-Llanos (2013)	More than 4 M tokens from 4,204 technical texts in Spanish	Only Part-of-Speech tagging
IxaMedGS Oronoz et al. (2015)	41,633 tokens from 75 clinical reports	2,362 diseases (490 negated and 40 speculated), 404 allergies (273 negated and 13 speculated), 1,191 drugs and 228 Adverse Drug Reactions (ADR) relations
MANTRA corpus Kors et al. (2015)	1,961 tokens and 100 texts written in Spanish from EMA 1,087 tokens and 100 texts from Medline Multilingual	5,530 annotations of UMLS semantic types and CUIs (756 in Spanish)
Spanish ADR Segura-Bedmar et al. (2015)	26,519 tokens and 397 texts related to mental health and schizophrenia extracted from Forum Clinic	187 Drugs and 636 adverse drug reactions (ADR)
Drug Semantics Moreno et al. (2017)	226,729 tokens and 30 texts extracted of summaries of product characteristics	724 Diseases, 657 Drugs, 557 Measurements, 66 Excipients, 62 Compositions, 45 Dose Forms, 42 Routes, 37 Medicaments, 31 Foods, and 20 Therapeutic Actions
IULA-SCRC Marimón et al. (2017)	3,194 sentences extracted from 300 electronic clinical records	7 Body parts, 14 Substances, 1,064 Findings, and 93 Procedures, 1,207 **Negations**
SpRadIE Cotik et al. (2017)	513 radiology reports	Anatomy (4,398), Measure (3,210), Finding (2,637), Texture (1,890), Measure Type (1,127), Location (722), Abbreviation (880), Temporal (35), Multiword (788); 9 relation types (10,987), **Negations (1,207)**, **Uncertainty** (109)
UHU-HUVR Cruz et al. (2017)	8,412 sentences from clinical documents	2,298 sentences annotated with **negation** cues, scopes, and events
BARR2 Intxaurrondo et al. (2018)	1,433 tokens from 3,563 report cases	9,552 annotations of acronyms, abbreviations, and expanded terms
DIANN (2018)	500 abstracts from papers in the biomedical domain	Disabilities
SPACCC Gonzalez-Aguirre et al. (2019)	396,988 tokens and 1,000 clinical cases from SciELO	3,009 Proteins (PharmaCoNER) Normalizable to SNOMED CT (4,398), Not- normalizable (50), Unclear (167) CODIESP: 18,483 ICD-10 codes
CANTEMIST Miranda-Escalada et al. (2020a)	1,051 Spanish clinical cases related to cancer	Tumor morphology mentions ICD-O-3 codes
CWLC Báez et al., (2020)	36,157 tokens and 1,912 sentences from referrals	9,029 entities (disease, body part, medication, symptoms, diagnostics, procedures, family member, abbreviation, result) attributes (385, 5 types) relations (284)
Nubes Lima et al. (2020)	7,019 electronic health records (29,682 sentences)	**Negation** (7,567 sentences) and **Uncertainty** (2,219 sentences)

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 7. Continued

Dataset	Details	Annotations
eHealth-KD Challenge Dataset (2021) Guijarro et al. (2021)	1,173 Spanish sentences in the medical domain from MedlinePlus	Entities (7,188), Roles (3,586) and 4 types of relations (2,339)
LivingNER Miranda-Escalada et al. (2022b)	2,000 Spanish clinical cases	Species Pathogens Foods NCBI codes
DisTEMIST Miranda-Escalada et al. (2022c)	1,000 Spanish clinical cases	Diseases Snomed-CT codes
SocialDisNER Sánchez et al., (2022)	10,000 health-associated tweets written in Spanish	Diseases

DIANN corpus available at: http://nlp.uned.es/diann/

Table 8. Spanish datasets in other domains

Dataset	Details	Annotations
UAM Spanish Treebank Sandoval and Salazar (2013)	1,501 sentences from newspapers	**Negation** cues and scopes (3,022)
SFU ReviewSP NEG Jiménez-Zafra et al., (2018)	9,455 sentences from product reviews	**Negation** cues, scopes, and events
NewsComm Taulé et al., (2021)	4,980 sentences of comments	Negation cues, scopes, events, and focus
T-MexNeg Bel-Enguix et al. (2021)	13,704 sentences from Mexican Spanish Tweets	**Negation** cues, scopes, and events

Likewise, multi-word uncertainty cues (UNC) frequency is slightly higher than the number of one-word uncertainty cues, but way shorter than uncertainty scopes (USCO), which are mostly long multi-word sequences also.

From Figure 6, an important additional fact can be claimed. Around 42.8% of negation references are out of the negation scope of the negated statement (40.7% in training, 44.7% in development, and 42.9% in testing). This evidence, which we illustrate in Figure 8, backs up our goal of creating a dataset and fine-tuning a model for OSR identification and linking. This would contribute to restoring the three essential components of a negated statement (i.e., OSR, negation cue, and negation scope).

3.3 Annotation process and guidelines

This section describes the dataset annotation process to augment NUBES into NeRUBioS. As mentioned above, we address negation references that have been so far left out by negation identification systems because they are out of the scope of the relevant negation cue (see example in Figure 1). We therefore follow Lima et al. (2020) to annotate OSRs and add them to NeRUBioS together with the annotations inherited from NUBES. This dataset was manually annotated by a computational linguist, native speaker of Spanish, and was thoroughly checked several times to ensure annotation consistency and terminology accuracy with the support of medical terminology

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

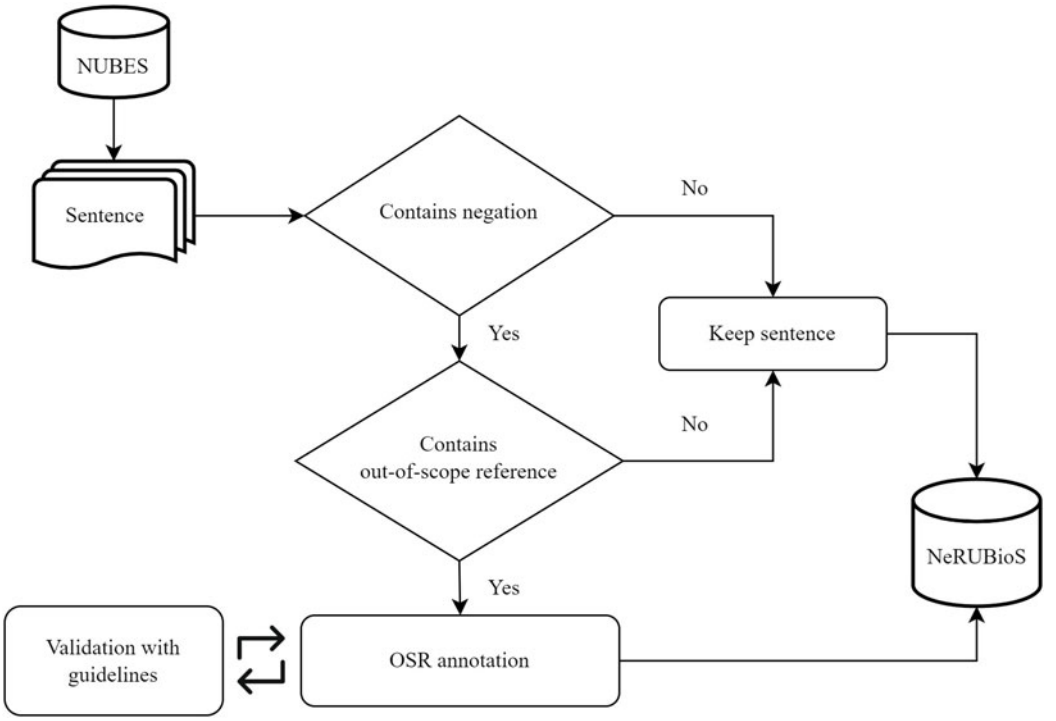

Figure 4. Augmenting NUBES into NeRUBioS.

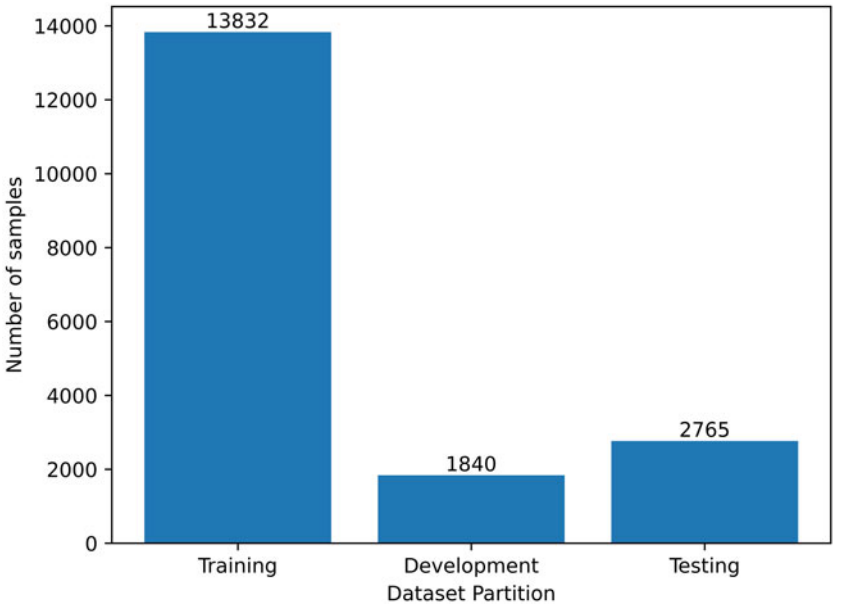

Figure 5. Number of samples across NeRUBioS dataset partitions.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 9. NeRUBioS general statistics

Feature/Label	OSR	Neg. Cue	Neg. Scope	Unc. Cue	Unc. Scope
Sentences with label	3,606	7,520	6,879	2,175	2,165
Multiword	2,636	711	6,261	951	2,073
Discontinuous	336	5	16	0	14
Sentences with multiple occurrences of label	226 (1.2%)	1,277 (6.93%)	1,171 (6.35%)	272 (1.47%)	282 (1.53%)
Average length	3.65 ± 3.21 min. 1 max. 25.7	1.08 ± 0.31 min. 1 max. 5.7	3.90 ± 3.62 min. 1 max. 42.3	1.41 ± 0.55 min. 1 max. 4.7	5.08 ± 4.68 min. 1 max. 35.3

Table 10. NeRUBioS tagset

Label	Meaning
B-NegREF	Beginning of OSR
I-NegREF	Inside of OSR
B-NEG	Beginning of negation
I-NEG	Inside of negation
B-NSCO	Beginning of negation Scope
I-NSCO	Inside of negation scope
B-UNC	Beginning of uncertainty
I-UNC	Inside of uncertainty
B-USCO	Beginning of uncertainty scope
I-USCO	Inside of uncertainty Scope
O	Outside (does not match any)

experts. Likewise, as we use Lima et al. (2020) methodology and annotations as an anchor to add OSR labels, we leverage and assume their inter-annotator agreement rate to ensure NeRUBioS's annotation reliability, which averages 85.7% for negation and uncertainty cues and scopes.

The main challenges posed by this task are:

(1) The semantic or syntactic role of OSRs in the sentences is unpredictable, which rules out any sort of rule-based assisted annotation.
(2) Most OSRs are multi-word sequences, which may include conjunctions and commas. For example:
   - *abdomen blando y depresible* (soft and depressible abdomen)
   - *molestia abdominal infraumbilical, difusa, constante* (diffuse, constant, infra umbilical abdominal discomfort)

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

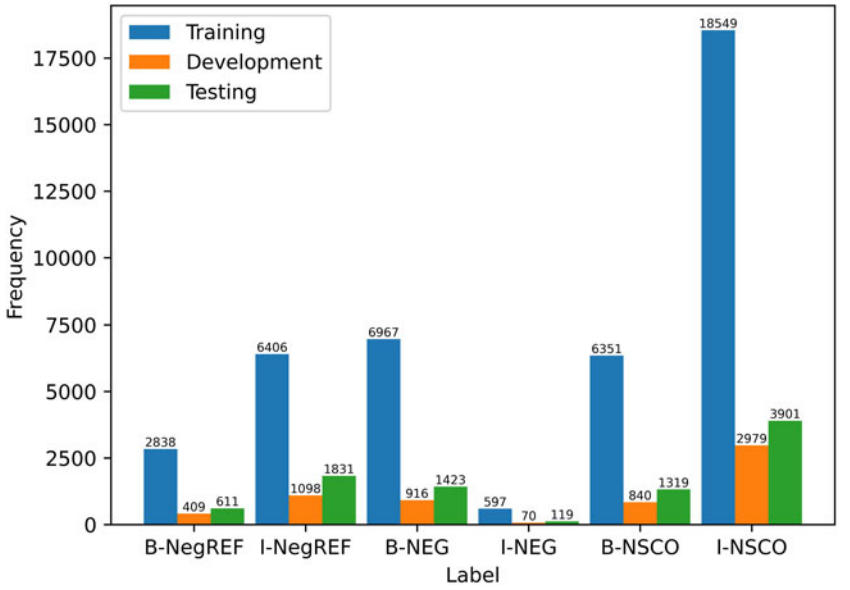

Figure 6. Negation label distribution.

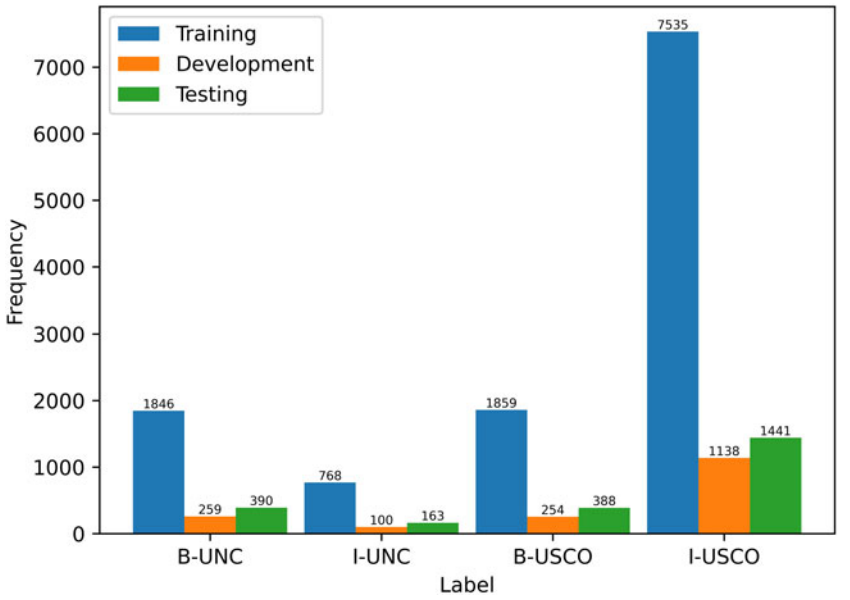

Figure 7. Uncertainty label distribution.

(30) OSRs may be discontinuous. For example, in the sentence below, the string *nódulos en CV* comes between two parts of an OSR (i.e., *Fibroinolaringoscopia* and *cierre*):

Fibroinolaringoscopia: *nódulos en CV*, *cierre incompleto*.

(4) OSRs and their negations are not always adjacent.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

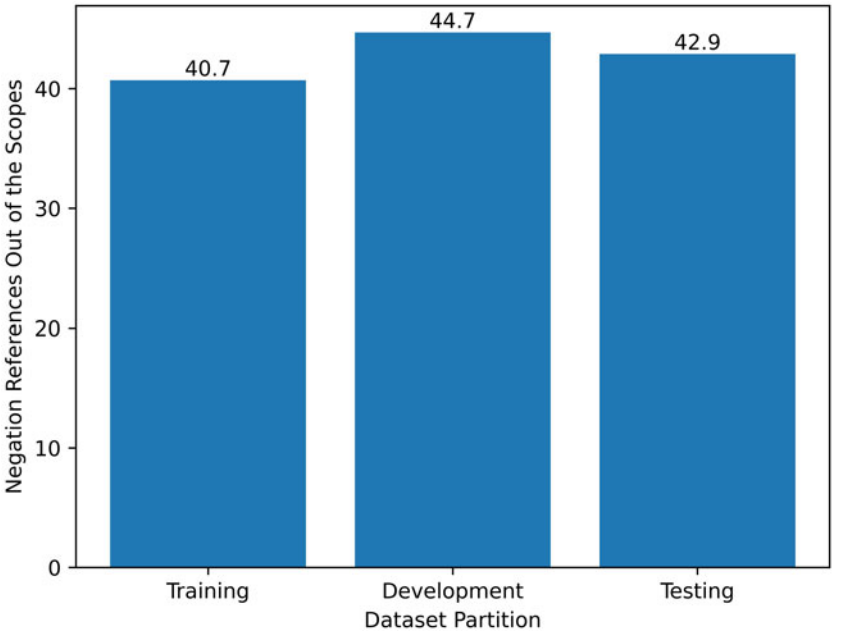

Figure 8. Out-of-scope negation references across partitions.

(5) There can be more than one valid OSR for a single negation in the same sentence. If this is the case and the OSRs are of different types, we call them *mixed* (see Table 11).

(6) OSRs may fall within the scope of another class, such as an uncertainty scope.

(7) There can be ambiguity when the same sentence includes two or more negations.

We found that the latter challenge barely represents 1.2% of the OSRs in the dataset, and, therefore, it was not addressed. The first six challenges were tackled by following the annotation guidelines below.

As a general rule, to delimit the boundaries of an OSR, we identify its head (mostly nouns) and extend it bidirectionally until the maximal syntactic unit is covered (e.g., noun phrase or verb phrase). That is, this unit must include its attributes or modifiers. Determiners such as definite articles, however, are annotated as "O," that is, outside of the OSR (e.g., *the*_O *admitted*_B-NegREF *patient*_I-NegREF). Likewise, during the OSR annotation process, when an OSR fell within an uncertainty scope, priority was given to the OSR annotation and uncertainty labels were removed.

With this general approach in mind, we manually annotated 13,193 tokens for a total of 3,858 OSRs in the dataset. As the annotation progressed, a pattern arose, namely, every OSR always comes before the negation, either adjacently or not. This is because, when the negation reference comes after the negation, it is generally captured by the negation scope. The pattern of a negated statement with OSR, therefore, is as follows:

Out-of-scope reference + *other possible items* + ***negation cue*** + *[negation scope]*

This pattern helped with the identification of categories the OSRs fall into. Roughly, these categories are diseases (findings), body parts (organs, tissues, etc.), types of tests or examinations, individuals or groups of people, medications, treatments, procedures, actions, and combinations of them (i.e., mixed OSRs). Table 11 lists bilingual examples of each category. The ellipsis (. . .)

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 11. Typology and examples of OSRs

OSR Category	Example of OSR + negation scope [spa-eng]
Disease	. . . [*cirrosis hepática*][***no*** *conocida previamente*]. . . . [cirrhosis of the liver][**not** previously known].
Organ/body part	[*labrum anterior y posterior*][***no*** *presentan alteraciones*]. [anterior and posterior labrum][show **no** alterations].
Test/examination	[*TAC abdominopélvico*] . . . [***sin*** *hallazgos*]. [Abdominopelvic CT] . . . [**no** findings].
Individual/group	[*paciente*] . . . [***a****sintomático*]. [patient] . . . [**a**symptomatic].
Medication	[*Digoxina*] . . . [***no*** *toma martes, jueves ni sábado*]. [Digoxin] . . . [**does not** take Tuesday, Thursday, or Saturday].
Treatment	. . . [*antibioterapia endovenosa*] . . . [***sin*** *incidencias*] . . . . . . [intravenous antibiotherapy] . . . [**without** incident] . . .
Procedures	. . . [*colocación de sonda de gastrostomía*] . . . [***sin*** *complicaciones*] . . . [gastrostomy tube placement] . . . [**without** complications].
Action	[*continuidad de cuidados*] . . . [***no*** *se podía realizar a través de atención primaria*] . . . [continuity of care] . . . [**could not** be provided through primary care] . . .
Expert's opinion	[*Desde el punto de vista urológico*] . . . [***sin*** *más actuación*]. [From the urological point of view] . . . [**without** further action].
Disease + Treatment	. . . [*Neumonía retrocardíaca*] . . . [*tto empírico con Cefotaximina y Levofloxacino 15 días*] [***sin*** *mejoría radiológica de la misma*].
	. . . [Retrocardiac pneumonia] . . . [empirical treatment with Cefotaxime and Levofloxacin was started for 15 days][**without** radiological improvement].
Individual + Action	. . . [*paciente camina*][***sin*** *ayuda*]. [Patient walks][**without** assistance].
Disease + Individual	. . . [*distorsión visual*] . . . [*paciente*][***incapaz*** *de precisar con claridad*] . . . . . . [visual distortion] . . . [patient][**unable to** pinpoint clearly] . . .
Procedure + Individual	. . . [*radiografía de tórax*] . . . [*paciente*][***no*** *presentaba condensación neumónica*] . . . . . . [chest X-ray] . . . [patient][showed **no** pneumonic condensation] . . .
Procedure + Disease	[*Intervenido*] . . . [*verrugas en la vejiga*] . . . [***sin*** *informes*]. [Operated] . . . [bladder warts][**no** reports].
Individual + Disease	[*paciente*] . . . [*lumbago*] . . . [***nula*** *mejoría*] . . . [Patient] . . . [back pain] . . . [with **no** improvement] . . .

between an OSR example and its negation indicates that they are not adjacent, and the plus (+) sign between categories means a mixed OSR.

3.4 Fine-tuning experiments

We used NeRUBioS to fine-tune a number of state-of-the-art Large Language Models (LLMs) based on BERT (Devlin et al., 2018). These models were retrained to tackle three tasks at a time using transfer learning, that is, negation detection, speculation detection, and OSR detection and linking with its negation scope. We report the results of five representative models that we utilized for the fine-tuning process:

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

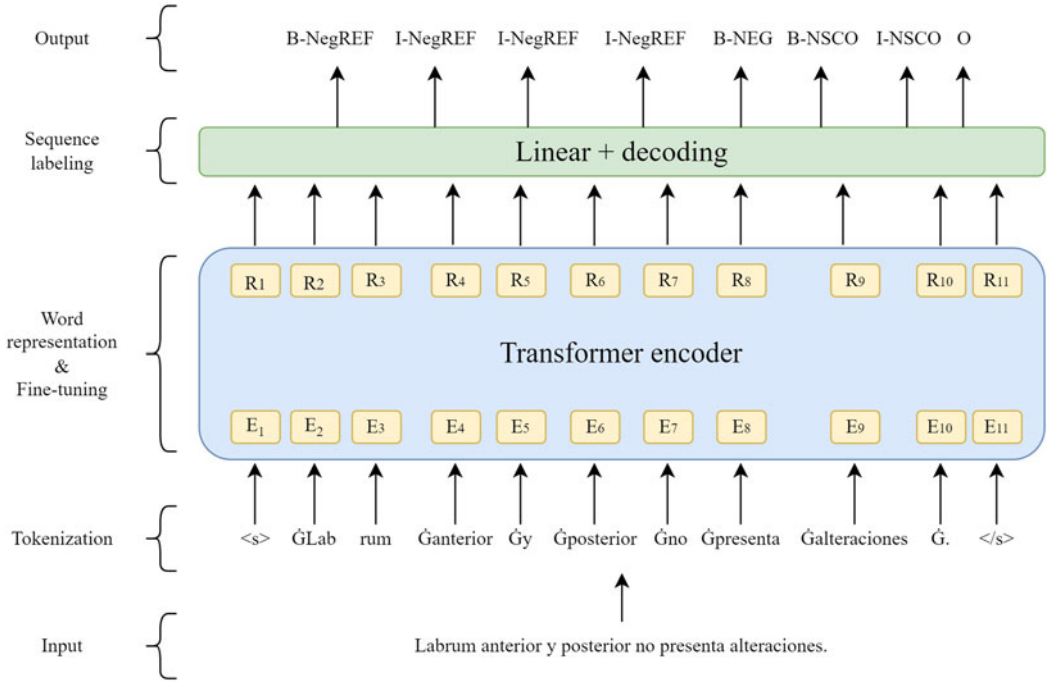

Figure 9. The model's architecture.

- RoBERTa (Carrino et al., 2021). The version of RoBERTa that we used is a pre-trained LLM in the medical field. The corpus used during its training consists of a collection of documents in the biomedical-clinical domain from several sources in Spanish.
- mBERT (Devlin et al., 2018). It is a version of BERT trained with large documents written in 104 different languages, including Spanish. mBERT was trained to predict both the next sentence and masked words.
- BETO (Cañete et al., 2020). BETO is a Spanish version of BERT trained from scratch on a big corpus. This model's size is similar to BERT-base's and was trained using the word masking technique.
- RoBERTa-BNE (Gutiérrez-Fandiño et al., 2022). A general-domain Spanish version of the RoBERTa architecture, pre-trained on a 570 GB corpus obtained from the Spanish National Library (BNE)
- XLM-RoBERTa (Conneau et al., 2020). A model pre-trained on the general-domain 2.4TB CommonCrawl Corpus in 100 languages.

In order to be able to jointly tackle these tasks as a token classification problem, each model's prediction layer was modified. Figure 9 shows the architecture we used to fine-tune the models. The architecture receives a sentence as input, carries out a subword tokenization process (RoBERTa's tokenization is shown), transforms the tokens to word embeddings encoding the token position in the sentence, and generates a powerful representation for each token by applying the multi-head self-attention process. Then, the representation of the input is passed to a linear layer which predicts a label for each token. Finally, a decoding process is carried out for a subword concatenation to obtain the output.

We followed the same training, development, and test partitions described above for the fine-tuning process with all the models. To find the best hyperparameter configuration, a grid search

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 12. Hyperparameters used during training

Hyperparameter	Value
Learning rate	2e-5
Training batch size	8
Evaluation batch size	8
Seed	42
Optimizer	Adam with betas = (0.9,0.999); epsilon = 1e-8
Learning rate scheduler type	Linear
Number of epochs	12
Loss function	cross-entropy loss
Activation function	GELU
Weight decay	0.1

for epochs (12, 7, 5, 3), learning rate (5e-7, 5e-5, 2e-7, 2e-5), and weight decay (0.01, 0.1) were used. 12, 2e-5, and 0.1 were the best values for epochs, learning rate, and weight decay, respectively (see Results below). The remaining hyperparameters were kept by default (see Table 12 below).

We used the transformer library and the models available at Hugging Face Hub. A Tesla A100 GPU with 27 GB of RAM memory was used for all the experiments. We recorded the computational cost during the fine-tuning process for RoBERTa, mBERT, and BETO, which was 80, 50, and 50 min respectively. The NeRUBioS dataset and the code implemented for this work are ready to be released at a public repository upon publication of this article.

4. Results and discussion

This work's overall goal is to automatically identify out-of-scope negation references and to link them to their respective negation markers and scopes in the sentence. In this section, we describe the obtained results for each individual task and then we evaluate OSR detection and linking together as well as all the tasks, including uncertainty detection. The values in this sections' tables are the average results of 12 epochs at the hyperparameter configuration described above. Likewise, the figures in this section show a plot of the F1 score for the five tested models at each of the 12 epochs in both the development and the testing partitions.

The results of the models have been evaluated using precision (Pr), recall (Rec), and F-score (F1). These metrics have a scale of 0.0 to 1.0, and are defined in equations (1), (2), and (3). Since the problem faced is treated as a token classification one, we used the average precision, the average recall, and the average F1 score. The average F1 score is the harmonic mean of precision and recall.

$$Precision = \frac{Number\ of\ tokens\ correctly\ classified}{Number\ of\ classified\ tokens} \tag{1}$$

$$Recall = \frac{Number\ of\ tokens\ correctly\ classified}{Number\ of\ tokens\ in\ the\ dataset} \tag{2}$$

$$F1score = \frac{2 * precision * recall}{precision + recall} \tag{3}$$

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 13. Results for OSR detection

	Out-of-Scope Reference					
	Development			Testing		
Model	Precision	Recall	F1score	Precision	Recall	F1score
RoBERTa	0.59 ± 0.02	0.59 ± 0.01	**0.59** ± 0.01	0.53 ± 0.04	0.59 ± 0.04	**0.56** ± 0.04
XLM RoBERTa	0.53 ± 0.04	0.55 ± 0.02	0.54 ± 0.03	0.53 ± 0.04	0.57 ± 0.05	0.55 ± 0.04
mBERT	0.53 ± 0.01	0.55 ± 0.01	0.54 ± 0.01	0.49 ± 0.05	0.55 ± 0.04	0.52 ± 0.04
BETO	0.54 ± 0.02	0.55 ± 0.03	0.54 ± 0.01	0.48 ± 0.05	0.55 ± 0.03	0.51 ± 0.03
RoBERTa BNE	0.51 ± 0.03	0.53 ± 0.02	0.52 ± 0.01	0.51 ± 0.03	0.55 ± 0.03	0.53 ± 0.03

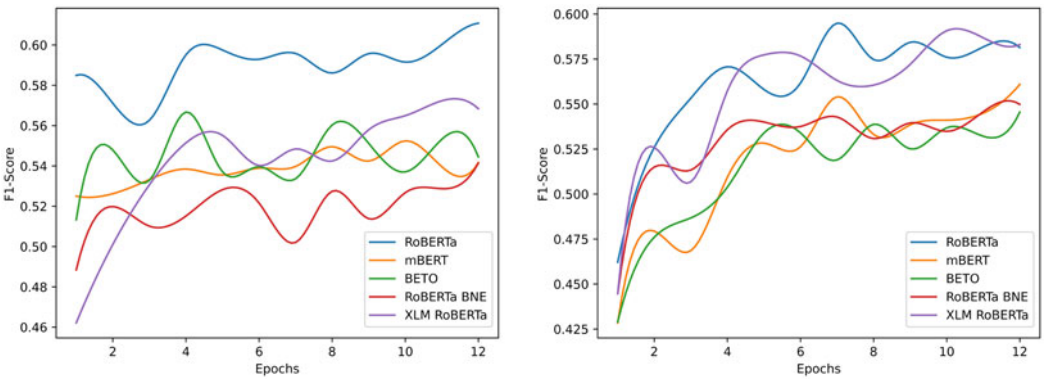

Figure 10. Model performance by epochs for OSR detection on the development (left) and testing (right) partitions.

4.1 OSR detection

RoBERTa was the best model for this task and reached F1 scores of 0.59 and 0.56 on the development and testing datasets, respectively (see Table 13 and Figure 10). To the best of our knowledge, this is the first time this problem has been addressed, which makes these results not only a novel contribution to the field but also a competitive baseline, given the complexity of the task.

As we treat the OSR identification problem as a token classification task, we do not measure whether the whole OSR was correctly labeled but rather whether each token in the OSR was correctly classified. The underlying philosophy of this methodology is that, in an information management system, a truncated OSR is more helpful than no OSR at all. However, in order to provide a deeper insight into the results in Table 13, we took a further step supported on the statistics shown in Table 9. Notice that OSRs' length ranges from 1 to 25 words with an average length of 3.65 words. These figures are relevant because it has been observed (see also Error Analysis below) that classification accuracy decreases as phrase length increases. Based on this assumption, we extrapolated the test F1 score in Table 13 to estimate F1 scores for OSRs shorter and longer than the average length 3.65. This may be helpful in the future to assess the usefulness of some pre- or post-processing to handle OSRs' length.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

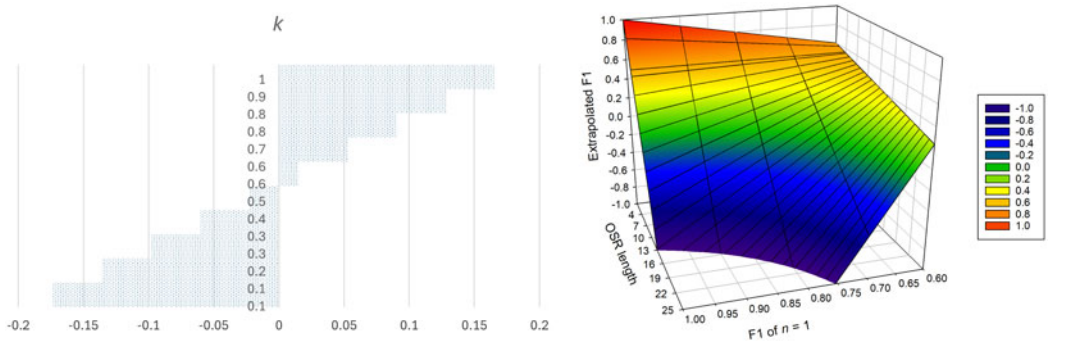

Figure 11. Obtained *k* values (left) and Extrapolated F1 scores (right).

We can use linear interpolation to estimate the interpolation for predicting the F1 score for OSRs of lengths incrementing from 1 to 25 based on the observed F1 score of 0.56 in the testing dataset for the average OSR length of 3.65. This interpolation assumes a linear relationship between the lengths of the OSRs and the F1 scores. Let's denote:

- $F1_{3.65} = 0.56$: The observed F1 score for OSRs of average length 3.65.
- $F1_1$: F1 score for OSRs of length 1. We will need this value to determine the rate of decrease.

The linear decay in F1 score as the OSR length increases is considered in the interpolation formula:

$$F1_n = F1_{3.65} + k * (n - 3.65)$$

where k is the rate of decrease in F1 score with respect to OSR length and $n - 3.65$ represents the deviation of the OSR length n from the average length 3.65.

We need to determine the value of k based on the observed F1 score for length 1 ($F1_1$). We can calculate it as follows:

$$k = \frac{F1_1 - F1_{3.65}}{3.65 - 1}$$

Once we have the value of k, we can use it in the interpolation formula above to estimate the F1 score for OSRs of different lengths.

In order to determine k, we need the F1 score of OSRs of length 1 ($F1_1$). We calculated k with a different $F1_1$ score each time starting with 0.1 up to 1 at intervals of 0.1. Figure 11 (left) shows that $F1_1$ scores below 0.6 yield negative k values, which makes sense if we consider that 0.56 is the reference score to start calculating the decay.

To visualize the way OSR length and $F1_1$s impact extrapolated F1 scores, we plot a mesh diagram (see Figure 11, right) including hypothetical $F1_1$ scores for OSRs of length 1 ($n = 1$) from 0.6 to 1, OSR lengths from 1 to 25, and the resulting extrapolated F1 scores. This model assumes that, the shorter the OSR, the better the token classifier's performance. The graph helps identifying the thresholds to avoid OSR lengths and $F1_1$ scores combinations that yield negative extrapolated F1 values. Factoring in this decay, the average extrapolated F1 score for OSRs of length 1 is 0.8, and for OSRs of length 2 is 0.71, while for lengths 4 and 5, the extrapolated F1 score is 0.53 and 0.44, respectively.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 14. Results for negation detection

	Dev						Test					
	Cue			Scope			Cue			Scope		
Model	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1
RoBERTa	0.95	0.97	**0.96**	0.87	0.91	**0.89**	0.95	0.97	**0.96**	0.87	0.91	**0.89**
	±	±	±	±	±	±	±	±	±	±	±	±
	0.01	0.002	0.003	0.01	0.01	0.01	0.01	0.004	0.01	0.02	0.003	0.01
RoBERTa	0.95	0.97	0.96	0.85	0.89	0.87	0.95	0.96	0.96	0.86	0.89	0.88
BNE	±	±	±	±	±	±	±	±	±	±	±	±
	0.009	0.005	0.005	0.02	0.01	0.02	0.006	0.004	0.004	0.02	0.008	0.01
XLM	0.94	0.97	0.95	0.85	0.90	0.87	0.94	0.96	0.95	0.86	0.9	0.88
RoBERTa	±	±	±	±	±	±	±	±	±	±	±	±
	0.02	0.005	0.01	0.04	0.02	0.03	0.01	0.004	0.007	0.02	0.01	0.02
mBERT	0.94	0.98	0.96	0.86	0.90	0.88	0.94	0.96	0.95	0.85	0.89	0.87
	±	±	±	±	±	±	±	±	±	±	±	±
	0.01	0.003	0.004	0.01	0.01	0.01	0.01	0.003	0.01	0.03	0.01	0.02
BETO	0.95	0.97	0.96	0.84	0.88	0.86	0.94	0.97	0.95	0.85	0.89	0.87
	±	±	±	±	±	±	±	±	±	±	±	±
	0.01	0.003	0.004	0.01	0.01	0.01	0.01	0.004	0.004	0.02	0.01	0.01

4.2 Negation cue and scope detection

Results for negation on predicted cues and scopes are shown in Table 14 and Figures 12 and 13. Our fine-tuned RoBERTa reaches outstanding scores at both tasks (F1 = 0.96 and F1 = 0.89). Moreover, it is also robust enough to maintain negation detection performance despite being fine-tuned on NeRUBioS, a dataset augmented with more classes to encode an additional layer of information.

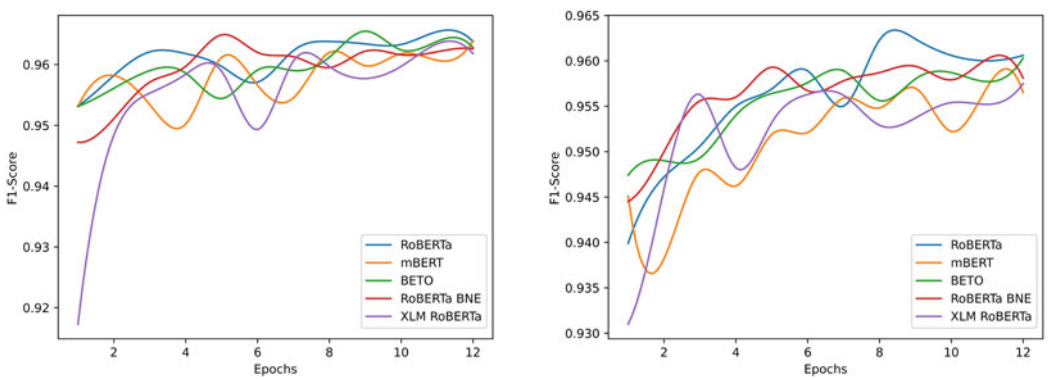

Figure 12. Model performance by epochs for negation cues on the development (left) and testing (right) partitions.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

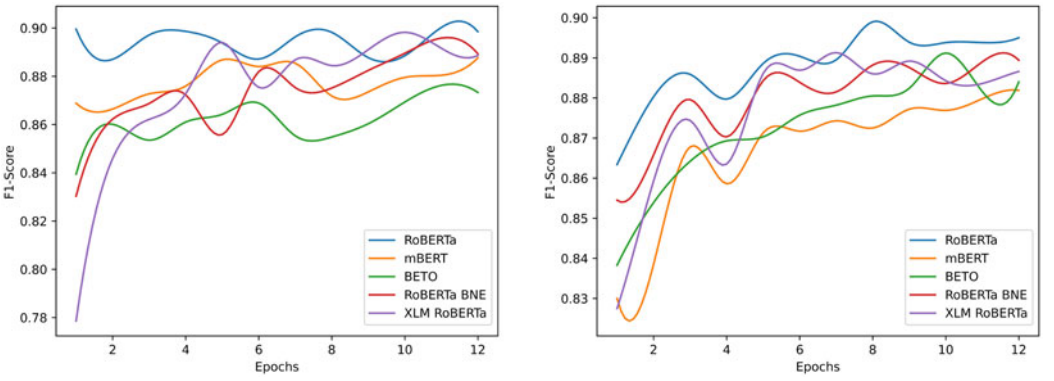

Figure 13. Model performance by epochs for negation scopes on the development (left) and testing (right) partitions.

Table 15. Results for uncertainty detection

	Dev						Test					
	Cue			Scope			Cue			Scope		
Model	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1
RoBERTa	0.83 ± 0.02	0.88 ± 0.01	0.85 ± 0.01	0.67 ± 0.03	0.75 ± 0.02	**0.71** ± 0.02	0.80 ± 0.03	0.87 ± 0.03	**0.84** ± 0.03	0.67 ± 0.03	0.77 ± 0.03	**0.71** ± 0.03
mBERT	0.83 ± 0.02	0.88 ± 0.01	**0.86** ± 0.01	0.65 ± 0.01	0.73 ± 0.01	0.69 ± 0.01	0.78 ± 0.02	0.84 ± 0.02	0.81 ± 0.02	0.64 ± 0.03	0.73 ± 0.03	0.68 ± 0.03
BETO	0.83 ± 0.02	0.87 ± 0.01	0.85 ± 0.01	0.62 ± 0.03	0.71 ± 0.02	0.66 ± 0.02	0.80 ± 0.02	0.87 ± 0.03	0.83 ± 0.02	0.65 ± 0.05	0.75 ± 0.03	0.70 ± 0.04
RoBERTa BNE	0.82 ± 0.01	0.85 ± 0.04	0.84 ± 0.02	0.63 ± 0.03	0.71 ± 0.04	0.66 ± 0.03	0.80 ± 0.02	0.85 ± 0.02	0.83 ± 0.01	0.65 ± 0.03	0.73 ± 0.02	0.69 ± 0.02
XLM RoBERTa	0.82 ± 0.04	0.87 ± 0.02	0.85 ± 0.03	0.60 ± 0.06	0.72 ± 0.02	0.66 ± 0.05	0.79 ± 0.03	0.87 ± 0.02	0.83 ± 0.02	0.62 ± 0.06	0.75 ± 0.03	0.68 ± 0.05

4.3 Uncertainty cue and scope detection

Results for uncertainty detection on predicted cues and scopes are less consistent (see Table 15 and Figures 14 and 15) and below-related works, though, as we show later. We think this may be explained by a number of cases of uncertainty scopes and OSRs that overlapped in the same text span in NeRUBioS. During the OSR annotation process, whenever an OSR fell within an uncertainty scope, priority was given to the OSR annotation and uncertainty labels were removed.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 16. Comparison with works doing negation and uncertainty detection using the NUBES dataset

	Neg.						Unc.					
	Cue			Scope			Cue			Scope		
Model	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1
Lima et al. (2020)	0.95	0.95	0.95	0.93	0.88	0.91	0.87	0.83	0.85	0.83	0.74	0.79
Solarte-Pabón et al. (2021b)	0.96	0.94	0.95	0.91	0.87	0.89	0.92	0.89	**0.90**	0.87	0.84	**0.85**
Pabón et al. (2022)	0.96	0.95	0.95	0.92	0.93	**0.92**	0.86	0.83	0.84	0.82	0.79	0.80
Our approach	0.95	0.97	**0.96**	0.87	0.91	0.89	0.80	0.87	0.84	0.67	0.77	0.71

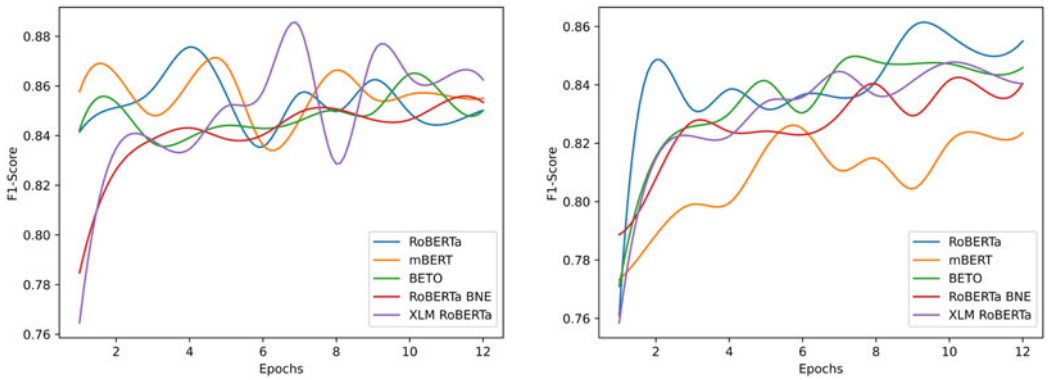

Figure 14. Model performance by epochs for uncertainty cues on the development (left) and testing (right) partitions.

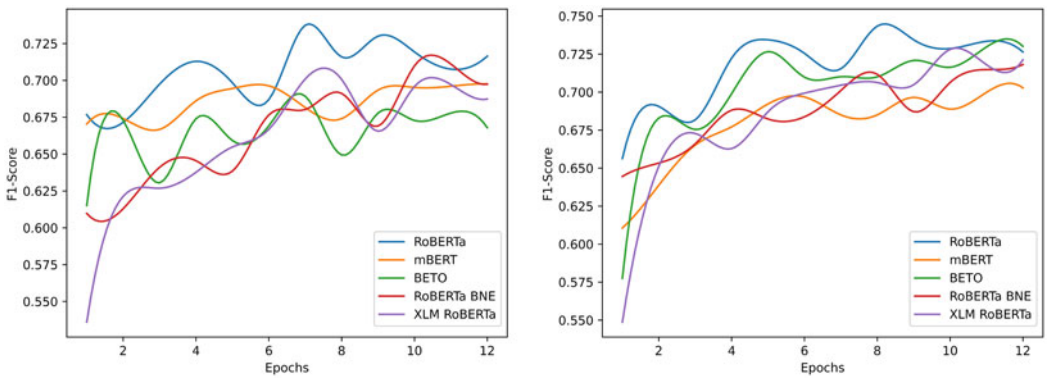

Figure 15. Model performance by epochs for uncertainty scopes on the development (left) and testing (right) partitions.

4.4 Negation and uncertainty detection results in perspective

Other works have reported very successful results for negation and uncertainty identification on the NUBES dataset. Table 16 and Figure 16 include these works as well as our own results with NeRUBioS so we can put them all in perspective. It is not clear, though, whether Solarte-Pabón et al. (2021b) used the same testing partition the other authors cited in the table used for these scores. The results show that fine-tuning our model with more (OSR) labels does not impact negation detection performance; actually, our model performs slightly better for negation cue

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 17. Results for the OSR detection and linking task

	Development			Testing		
Model	Precision	Recall	F1score	Precision	Recall	F1score
RoBERTa	0.84 ± 0.01	0.87 ± 0.004	**0.85** ± 0.003	0.85 ± 0.01	0.87 ± 0.01	**0.86** ± 0.01
RoBERTa-BNE	0.83 ± 0.01	0.85 ± 0.02	0.84 ± 0.02	0.84 ± 0.01	0.85 ± 0.01	0.85 ± 0.01
XLM-RoBERTa	0.82 ± 0.02	0.86 ± 0.01	0.84 ± 0.02	0.84 ± 0.01	0.86 ± 0.01	0.85 ± 0.01
mBERT	0.82 ± 0.01	0.85 ± 0.01	0.84 ± 0.01	0.81 ± 0.03	0.85 ± 0.01	0.83 ± 0.02
BETO	0.83 ± 0.01	0.85 ± 0.01	0.84 ± 0.01	0.81 ± 0.01	0.86 ± 0.01	0.83 ± 0.01

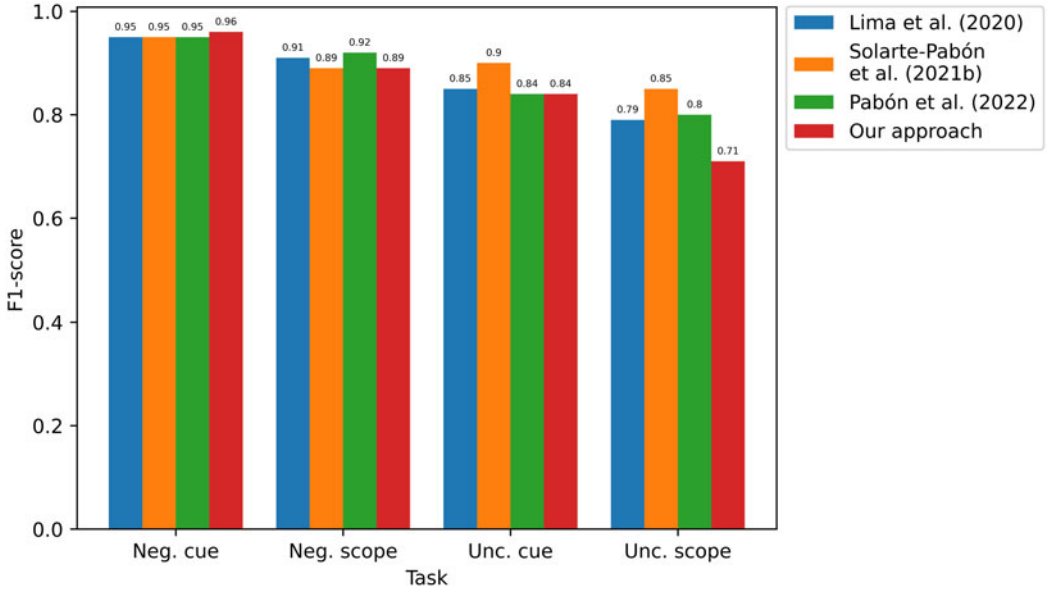

Figure 16. Comparison with state-of-the-art models.

detection. Uncertainty detection, however, was compromised to some extent due to a number of uncertainty labels being replaced with OSR labels, as we explained above.

4.5 OSR detection and linking

This task involves OSR detection, negation cue detection, and negation scope detection. While this task has been treated here like a token classification problem, it can be seen as one of relation extraction. This is so because most OSRs and their negation are in the same sentence. Therefore, once the OSR and its negation have been identified, their relation has likewise been established. Table 17 and Figure 17 show very competitive F-scores with RoBERTa in both partitions (0.84 and 0.86) for this novel task.

4.6 The joint task: OSR detection and linking + uncertainty detection

This joint task encompasses the five tasks assessed in the previous sub-sections, that is, OSR detection, negation cue detection, negation scope detection, uncertainty cue detection, and uncertainty

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 18. Overall results for the OSR detection and linking task + uncertainty detection

	Development			Testing		
Model	Precision	Recall	F1score	Precision	Recall	F1score
RoBERTa	0.83 ± 0.01	0.86 ± 0.003	**0.84** ± 0.01	0.81 ± 0.02	0.86 ± 0.01	**0.83** ± 0.01
mBERT	0.81 ± 0.01	0.85 ± 0.004	0.83 ± 0.01	0.79 ± 0.03	0.84 ± 0.01	0.81 ± 0.02
BETO	0.80 ± 0.01	0.84 ± 0.01	0.82 ± 0.01	0.79 ± 0.03	0.85 ± 0.01	0.82 ± 0.02
RoBERTa-BNE	0.80 ± 0.02	0.83 ± 0.01	0.82 ± 0.01	0.81 ± 0.01	0.84 ± 0.01	0.82 ± 0.01
XLM-RoBERTa	0.79 ± 0.04	0.84 ± 0.01	0.82 ± 0.03	0.78 ± 0.03	0.85 ± 0.01	0.82 ± 0.02

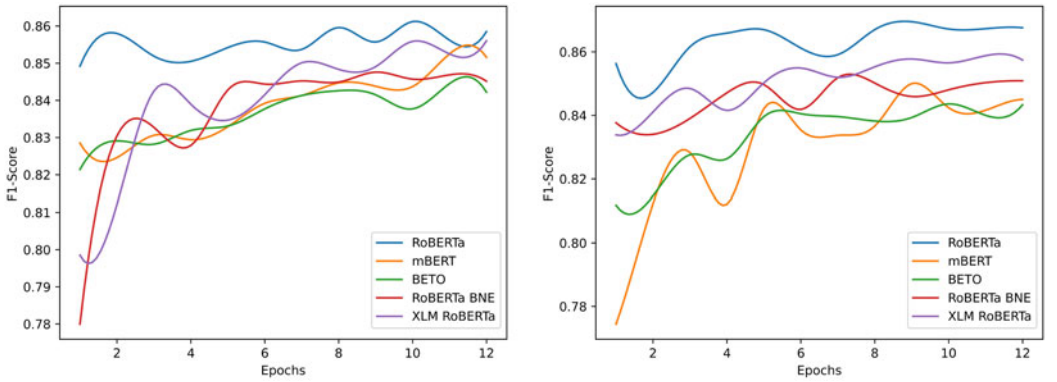

Figure 17. Model performance by epochs for OSR detection and linking task on the development (left) and testing (right) partitions.

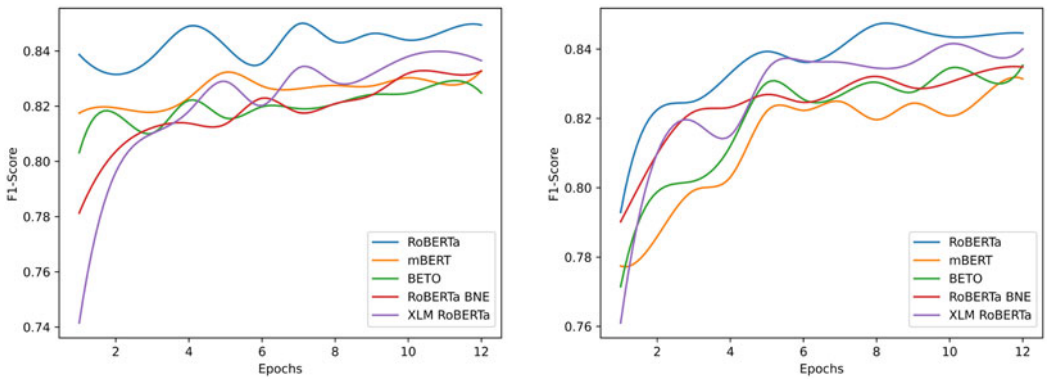

Figure 18. Model performance by epochs for OSR detection and linking + uncertainty detection on the development (left) and testing (right) partitions.

scope detection. Table 18 and Figure 18 show a competitive performance for the joint task, although there is variance in these scores brought in by both the high scores of negation detection and the not-so-high OSR detection scores.

With regard to the results in Table 18, RoBERTa certainly was the best model. However, mBERT's performance is remarkable. It outperforms the other models if its multilingual capacity is considered. mBERT can solve the same task in several languages with promising results using the zero-shot approach, as shown in the machine-translated examples in Table 19.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Table 19. Cross-lingual inference with mBERT

Original example in Spanish	Cross-lingual zero-shot inference with mBERT
Paciente con los antecedentes reseñados que ingresa por cuadro/NegREF de escasas horas de evolución consistente en exacerbación de su temblor habitual/NegREF, que parece/UNC haberse hecho generalizado/USCO y cuya descripción/NegREF es ***incapaz/NEG de/NEG*** *precisar/NSCO.*	**English:** Patient with the aforementioned history is admitted for a few hours of evolution consisting of exacerbation of/NegREF his usual tremor/NegREF, which seems/UNC to/UNC have become generalized/USCO and whose description/NegREF is **unable/NEG** to/NSCO specify/NSCO. **French:** Patient aux antécédents précités admis pour des symptômes de quelques heures d'évolution consistant en une exacerbation de son tremblement habituel/NegREF, qui semble/UNC s'être généralisé/USCO et dont il **n'/NEG est/NSCO pas/NEG** en mesure de préciser la description/NSCO.

4.7 Error analysis

We carried out an exhaustive error analysis in OSR detection and linking and have grouped errors in categories as detailed below. In the provided examples, gold and predicted OSRs are in bold while negation cues and their scopes are underlined.

- Discontinuous OSRs: Some discontinuous OSRs are not detected in their entirety. This type of OSR is very challenging for the model due to the occurrence of other items interrupting an OSR:

  Gold
  Paciente con los antecedentes reseñados queue ingresa por ***episodio de pérdida de conciencia*** *mientras se encontraba sentada, que se inicia* ***con sensación de mareo con prodromos de visión borrosa*** *sin otros síntomas. . .*

  Predicted
  con sensación de mareo con prodromos de visión borrosa *sin otros síntomas. . .*

  Like the example shows, the model tends to correctly predict the part that is closer to the negation cue. This was expected since there are many more samples of continuous OSRs in the dataset. Additionally, the first or last part of discontinuous OSRs may feature an infrequent syntactic pattern, which adds to the already challenging identification task.

- Mixed OSRs: Some mixed OSRs are not detected in its entirety, as in the example below where the OSR is made up of a disease (*cáncer de pulmón*) and a treatment (*tratado con quimioterapia*):

  Gold
  Cáncer de pulmón *detectado hace 8 semanas después de consulta por signos de dolor lumbar fue* ***tratado con quimioterapia*** *presentando nula mejoría.*

  Predicted
  tratado con quimioterapia *presentando nula mejoría.*

  Mixed OSRs generally include some sort of complementary information. Therefore, in most cases, the model can identify the part of the OSR corresponding to one category

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

of the OSR, but it struggles to extract the complementary part. Apparently, the myriad of patterns in mixed OSR causes the errors in this category.

- Long OSRs: Long OSRs are frequently not detected in their entirety. In many cases, the model can identify OSR chunks with a complete semantic sense, as in this example, but it is not always the complete OSRs:

  Gold
  Paladar asimétrico con desviación de uvula a la derecha, hiperemico*, no abombado.*

  Predicted
  Paladar *asimétrico con desviación de uvula a la derecha, hiperemico, no abombado.*

  OSRs in NeRUBioS were tagged by identifying their head (mostly nouns) and extending it bidirectionally until the maximal syntactic unit is reached. This sometimes produces very long OSRs, which can include punctuation marks, especially commas, and non-content words. This type of OSR is a challenge due to the high variety in their syntactic patterns.

- Tokenization of numbers: OSRs are sometimes truncated when numbers occur.

  Gold
  Micralax cánulas rectal *si más de 48 horas sin deposiciones.*

  Predicted
  de 48 horas sin deposiciones.

  Numbers are highly frequent in clinical documents. Due to the tokenization process carried out by the transformer architecture, some numbers are split causing the truncation of OSRs that include a number.

Overall, most of the strings resulting from the types of errors described above are truncations, missing OSR parts, or extended OSRs. As OSRs in NeRUBios are not categorized according to this error typology, it was not possible to quantify the number of errors in each category. However, all these categories together significantly impact the model's recall and precision, particularly because most of the OSRs are multi-word sequences. On the other hand, a qualitative analysis shows that, when the exact match constraint is relaxed, our approach can still identify useful OSR chunks and link them with their respective negation scopes. The number and usefulness of this partial matching cases, however, are not reflected on the evaluation metrics above, since we assess each model's exact predictions rather than any form of fuzzy string matching.

5. Conclusions

This work addressed the phenomenon of OSRs in negated statements in clinical documents in Spanish. OSRs are crucial to the integral meaning of a negated statement and they have been systematically left out by negation detection systems so far. Our survey of the literature up to date reveals that *(1)* this is the first time the issue has been tackled; *(2)* related issues such as negation and uncertainty/speculation detection have been tackled with four distinct approaches, that is, rules, classical machine learning, hybrid methods, and deep learning. These approaches can be seen as methodological generations given their appearance in chronological order with some overlapping between generations; *3)* the early works in each generation reached high performance as they initially tackled more formal features, but, as the modeling of the problem unveiled higher level, more complex features, each generation's performance as a whole seems to decrease over time, with the exception of deep-learning models.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

In order for the OSR task to be approached with transfer learning using deep-learning models, a dataset with annotated OSRs was required. Therefore, we manually augmented NUBES (Lima et al., 2020) into NeRUBioS, the first negation and uncertainty clinical Spanish dataset with OSR annotations. For this, a protocol was defined and followed. The annotation process unveiled seven challenges posed by OSRs, but it also allowed to determine the overall pattern of negated statements. OSRs fall into nine categories and account for 42.8% of negation references in the dataset, which makes it a very relevant problem. Using NeRUBioS, we fine-tuned five BERT-based models and used transfer learning to jointly identify negation scopes and their respective OSRs as well as uncertainty. Our best model achieves state-of-the-art performance in negation detection while also establishing a competitive baseline for OSR identification (Macro F1 = 0.56) and linking (Macro F1 = 0.86). Moreover, an extrapolation of these results to OSRs of shorter lengths suggests that the F1 score for this task may go up to 0.71 for two-word OSRs and to 0.80 for one-word OSRs. The results suggest that ORS identification may be more challenging for the models than negation and uncertainty detection. An analysis of errors and results confirms that the tested models struggle with some of the OSR challenges outlined in the methodology above plus other unforeseen features like dates and numbers. However, a qualitative assessment still shows very useful hits for OSR detection when the exact match constraint is relaxed a bit. Uncertainty detection performance, on the other hand, was impacted by the overlapping of OSRs and uncertainty scopes in some instances.

For future work, we plan to address the models' limitations, such as distant references, by combining deep-learning architectures to enhance OSR detection and linking results. This combination can also be in the form of a pipeline of models trained for various related tasks. For example, many OSRs fall in the categories of diseases and species, for which a myriad of named-entity recognition (NER) systems has been proposed. The model itself or an architecture enriched with some sort of layer with NER knowledge might help the identification of challenging cases such as discontinuous, mixed, and long OSRs. Additionally, while this work keeps the dataset partitions' size consistent with Lima et al. (2020) to allow for performance comparison, it will be interesting to balance or optimize the development and testing partitions size and see whether this, in combination with batch size, impacts the models' performance. Future iterations of NeRUBioS will also benefit from a dedicated measurement of inter-annotator agreement for OSRs. Likewise, we are adding post-processing capabilities to our system, which has proven useful in previous works in the medical field (Tamayo et al. (2022b); Tamayo et al. (2022c)). This may be particularly useful to solve number splits and, more importantly, to restore uncertainty tags removed during the OSR annotation process and boost uncertainty identification performance. This seems feasible since uncertainty scopes may overlap with OSRs but uncertainty cues rarely do; therefore, cue labels might be used as anchors for uncertainty scope restoration in a post-processing stage.

References

Agarwal S. and Yu H. (2010). Biomedical negation scope detection with conditional random fields. *Journal of the American Medical Informatics Association* **17**(6), 696–701.

Apostolova E., Tomuro N. and Demner-Fushman D. (2011). *Automatic extraction of lexico-syntactic patterns for detection of negation and speculation scopes*. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 283–287.

Attardi G., Cozza V. and Sartiano D. (2015). Detecting the scope of negations in clinical notes. CLiC it, 14.

Báez P., Villena F., Rojas M., Durán M. and Dunstan J. (2020). *The Chilean Waiting List Corpus: a new resource for clinical named entity recognition in Spanish*. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pp. 291–300.

Ballesteros M., Francisco V., Díaz A., Herrera J. and Gervás P. (2012). *Inferring the scope of negation in biomedical documents*. In Proceedings, Part I 13, *Computational Linguistics and Intelligent Text Processing: 13th International Conference, CICLing 2012*, New Delhi, India, Proceedings, Part I 13, March 11-17, 2012,: Springer Berlin Heidelberg, pp. 363–375.

Bel-Enguix G., Gómez-Adorno H., Pimentel A., Ojeda-Trueba S. L. and Aguilar-Vizuet B. (2021). Negation detection on mexican spanish tweets: The t-mexneg corpus. *Applied Sciences* **11**(9), 3880.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Beltrán J. and González M. (2019). Detection of negation Cues in Spanish: The CLiC-Neg system. In *IberLEF@ SEPLN*, pp. 352–360.

Cañete J., **Chaperon G.**, **Fuentes R.**, **Ho J. H.**, **Kang H. and Pérez J.** (2020). Spanish pre-trained bert model and evaluation data. *Pml4dc at ICLR* **2020**(2020), 1–10.

Carrino C. P., **Armengol-Estapé J.**, **Gutiérrez-Fandiño A.**, **Llop-Palao J.**, **Pàmies M.**, **Gonzalez-Agirre A. and Villegas M.** (2021). Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, arXiv preprint arXiv: 2109.03570.

Casillas A., **Pérez A.**, **Oronoz M.**, **Gojenola K. and Santiso S.** (2016). Learning to extract adverse drug reaction events from electronic health records in Spanish. *Expert Systems with Applications* **61**, 235–245.

Chapman W. W., **Bridewell W.**, **Hanbury P.**, **Cooper G. F. and Buchanan B. G.** (2001). *Evaluation of negation phrases in narrative clinical reports*. In *Proceedings of the AMIA Symposium, p. 105*, American Medical Informatics Association.

Conneau A., **Khandelwal K.**, **Goyal N.**, **Chaudhary V.**, **Wenzek G.**, **Guzmán F.**, **Grave E.**, **Ott M.**, **Zettlemoyer L. and Stoyanov V.** (2020, Unsupervised Cross-lingual Representation Learning at Scale. *arXiv [Cs.CL]*.

Cotik V., **Filippo D.**, **Roller R.**, **Uszkoreit H. and Xu F.** (2017). Annotation of entities and relations in spanish radiology reports. In *RANLP*, pp. 177–184.

Cotik V., **Stricker V.**, **Vivaldi J. and Rodríguez Hontoria H.** (2016). *Syntactic methods for negation detection in radiology reports in Spanish*. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing, BioNLP 2016*: Berlin, Germany, August 12, 2016, Association for Computational Linguistics, pp. 156–165.

Councill I. G., **McDonald R. and Velikovich L.** (2010). *What's great and what's not: learning to classify the scope of negation for improved sentiment analysis*. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, Association for Computational Linguistics, pp. 51–59.

Cruz Diaz N. P., **Mana López M. J.**, **Vázquez J. M.**, **lvarez Á. and V. P.** (2012). A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the American Society for Information Science and Technology* **63**(7), 1398–1410.

Cruz N. P., **Taboada M. and Mitkov R.** (2016). A machine-learning approach to negation and speculation detection for sentiment analysis. *Journal of the Association for Information Science and Technology* **67**(9), 2118–2136.

Cruz N., **Morante R.**, **Maña-López M. J.**, **Mata-Vázquez J. and Parra-Calderón C. L.** (2017). *Annotating Negation in Spanish Clinical Texts*. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles; Association for Computational Linguistics (ACL)*, Stroudsburg, PA, USA, pp. 53–58.

Dalianis H. (2018). *Clinical Text Mining: Secondary use of Electronic Patient Records*. Springer Nature.

Deléger L. and Grouin C. (2012). *Detecting negation of medical problems in French clinical notes*. In *Proceedings of the 2nd ACM Sighit International Health Informatics Symposium*, pp. 697–702.

Devlin J., **Chang M. W.**, **Lee K. and Toutanova K.** (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805.

Domínguez-Mas L., **Ronzano F. and Furlong L. I.** (2019). *Supervised learning approaches to detect negation cues in Spanish reviews*. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), CEUR Workshop Proceedings*, Bilbao, Spain: CEURWS.

Fabregat H., **Araujo L. and Martínez J.** (2019). Deep learning approach for negation trigger and scope recognition. *Procesamiento De Lenguaje Natural* **62**, 37–44.

Fabregat H., **Duque A.**, **Araujo L. and Martinez-Romo J.** (2021). LSI_UNED at CLEF eHealth2021: exploring the effects of transfer learning in negation detection and entity recognition in clinical texts. In *CLEF eHealth 2021. CLEF. 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*.

Fabregat H., **Duque A.**, **ínez-Romo J. and Araujo L.** (2019). *Extending a Deep Learning approach for Negation Cues Detection in Spanish*. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings*, CEUR-WS, Bilbao, Spain.

Fabregat H., **Martínez-Romo J. and Araujo L.** (2018). *Deep Learning Approach for Negation Cues Detection in Spanish at NEGES 2018*. In *En Proceedings of NEGES 2018: Workshop on Negation in Spanish*, Vol. 2174, pp. 43–48.

Fancellu F., **Lopez A. and Webber B.** (2016). *Neural networks for negation scope detection*. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers)*, pp. 495–504.

Fei H., **Ren Y. and Ji D.** (2020). Negation and speculation scope detection using recursive neural conditional random fields. *Neurocomputing* **374**, 22–29.

Fujikawa K., **Seki K. and Uehara K.** (2013). NegFinder: A web service for identifying negation signals and their scopes. *Information and Media Technologies* **8**(3), 884–889.

García-Largo M. and Segura-Bedmar I. (2021). Extracting information from radiology reports by Natural Language Processing and Deep Learning. In *CLEF (Working Notes)*, pp. 804–817.

Gindl S., **Kaiser K. and Miksch S.** (2008). Syntactical negation detection in clinical practice guidelines. *Studies in Health Technology and Informatics* **136**, 187–92.

Giudice V. (2019). *Aspie96 at NEGES: Negation Cues Detection in Spanish with Character-Level Convolutional RNN and Tokenization*. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings*, CEUR-WS, Bilbao, Spain.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Gkotsis G., **Velupillai S.**, **Oellrich A.**, **Dean H.**, **Liakata M. and Dutta R.** (2016). *Don't let notes be misunderstood: A negation detection method for assessing risk of suicide in mental health records*. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pp. 95–105.

Gonzalez-Aguirre A., **Marimon M.**, **Intxaurrondo A.**, **Rabal O.**, **Villegas M. and Krallinger M.** (2019). *PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track*. In *Proceedings of the 5th workshop on BioNLP open shared tasks Hong Kong*, China, 4 Nov. 2019, pp. 1–10.

Goryachev S., **Sordo M.**, **Zeng Q. T. and Ngo L.** (2006). Implementation and evaluation of four different methods of negation detection, DSG, Technical report, pp. 2826–2831.

Guijarro A., **Guillena R.**, **Morffis A.**, **Velarde S.**, **Gutiérrez Y. and Cruz Y. A.** (2021). Overview of the eHealth knowledge discovery challenge at iberLEF 2021. *Procesamiento del lenguaje natural* **67**, 233–242.

Gutiérrez-Fandiño A., **Armengol-Estapé J.**, **Pàmies M.**, **Llop-Palao J.**, **Silveira-Ocampo J.**, **Pio Carrino C.**, **Armentano-Oller C.**, **Rodriguez-Penagos C.**, **Gonzalez-Agirre A. and Villegas M.** (2022). MarIA: Spanish language models. *Procesamiento Del Lenguaje Natural* **68**, 39–60.

Gyawali B. and Solorio T. (2012). *UABCoRAL: A preliminary study for resolving the scope of negation*. In ∗ *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 275–281.

Harkema H., **Dowling J. N.**, **Thornblade T. and Chapman W. W.** (2009). ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics* **42**(5), 839–851.

Hartmann M. and Søgaard A. (2021). *Multilingual negation scope resolution for clinical text*. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pp. 7–18.

Huang Y. and Lowe H. J. (2007). A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association* **14**(3), 304–311.

Intxaurrondo A., **de la Torre J. C.**, **Rodríguez-Betanco H.**, **Marimon M.**, **Lopez-Martin J. A.**, **Gonzalez-Agirre A.**, **Santamaria J.**, **Villegas M. and Krallinger M.** (2018). *Resources, guidelines and annotations for the recognition, definition resolution and concept normalization of Spanish clinical abbreviations: the BARR2 corpus*. In *Proceedings of SEPLN*, pp. 1–9.

Jiménez-Zafra S. M., **Morante R.**, **Martín-Valdivia M. T. and Lopez L. A. U.** (2020). Corpora annotated with negation: An overview. *Computational Linguistics* **46**(1), 1–52.

Jiménez-Zafra S. M., **Morante R.**, **Blanco E.**, **Martín-Valdivia M. T. and López L. A. U.** (2020). *Detecting negation cues and scopes in Spanish*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6902–6911.

Jiménez-Zafra S. M., **Díaz N. P. C.**, **Morante R. and Martín-Valdivia M. T.** (2019). Neges 2018: workshop on negation in spanish. *Procesamiento Del Lenguaje Natural* **62**, 21–28

Jiménez-Zafra S. M., **Taulé M.**, **Martín-Valdivia M. T.**, **Urena-López L. A. and Martí M. A.** (2018). SFU review SP-NEG: a spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. *Language Resources and Evaluation* **52**, 533–569.

Khandelwal A. and Sawant S. (2019). *NegBERT: A Transfer Learning Approach for Negation Detection and Scope Resolution*. In *Proceedings of the LREC 2020-12th International Conference on Language Resources and Evaluation*, Marseille, France, 11 November 2019, pp. 5739–5748.

Kors J. A., **Clematide S.**, **Akhondi S. A.**, **Van Mulligen E. M. and Rebholz-Schuhmann D.** (2015). A multilingual gold-standard corpus for biomedical concept recognition: The Mantra GSC. *Journal of the American Medical Informatics Association* **22**(5), 948–956.

Lazib L., **Zhao Y.**, **Qin B. and Liu T.** (2019). Negation scope detection with recurrent neural networks models in review texts. *International Journal of High Performance Computing and Networking* **13**(2), 211–221.

Lima S., **Pérez N.**, **Cuadros M. and Rigau G.** (2020). *NUBes: A corpus of negation and uncertainty in Spanish clinical texts*. In *Proceedings of the 12th LREC*, Marseille, France, pp. 5772–5781, 2020-05-11

Loharja H., **Padró L. and Turmo Borras J.** (2018). *Negation cues detection using CRF on Spanish product review texts*, proceedings book, *NEGES 2018: Workshop on Negation in Spanish: Seville*, Spain, Proceedings Book, September 19-21, pp. 49–54.

López-Úbeda P., **Díaz-Galiano M. C.**, **López L. A. U. and Martín-Valdivia M. T.** (2021). Pre-trained language models to extract information from radiological reports. In *CLEF (Working Notes)*, pp. 794–803.

Mahany A., **Khaled H.**, **Elmitwally N. S.**, **Aljohani N. and Ghoniemy S.** (2022). Negation and speculation in NLP: A survey, corpora. *Methods, and Applications. Applied Sciences* **12**(10), 5209.

Mahany A., **Khaled H. and Ghoniemy S.** (2023). Arabic negation and speculation scope detection: a transformer-based approach. *International Journal of Intelligent Computing and Information Sciences* **23**(1), 29–40.

Marimón M., **Vivaldi J. and Bel N.** (2017). Annotation of negation in the IULA Spanish clinical record corpus. In *Proceedings of SemBEaR* **2017**, 43–52.

Martí M., **Taulé M.**, **Nofre M.**, **Marsó L.**, **Martín-Valdivia M. and Jiménez-Zafra S.** (2016). La negación en español: análisis y tipología de patrones de negación. *Procesamiento del Lenguaje Natural* **57**, 41–48.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Mehrabi S., Krishnan A., Sohn S., Roch A. M., Schmidt H., Kesterson J., Beesley C., Dexter P., Schmidt C. and Palakal M. (2015). DEEPEN: a negation detection system for clinical text incorporating dependency relation into NegEx. *Journal of Biomedical Informatics* **54**, 213–219.

Miranda-Escalada A., Farré E. and Krallinger M. (2020). Named entity recognition, concept Normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *IberLEF@ SEPLN*, pp. 303–323.

Miranda-Escalada A., Farré-Maduell E., Lima-López S., Estrada D., Gascó L. and Krallinger M. (2022). Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of the LivingNER shared task and resources. *Procesamiento del Lenguaje Natural* **69**, 241–253.

Miranda-Escalada A., Gascó L., Lima-López S., Farré-Maduell E., Estrada D., Nentidis A., Krithara A., Katsimpras G., Paliouras G. and Krallinger M. (2022). *Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources*. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings.*

Morante R. and Blanco E. (2012). * *sem 2012 shared task: Resolving the scope and focus of negation*. In * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 265–274.

Morante R. and Blanco E. (2021). Recent advances in processing negation. *Natural Language Engineering* **27**(2), 121–130.

Morante R. and Daelemans W. (2009). *A metalearning approach to processing the scope of negation*. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pp. 21–29.

Morante R., Liekens A. and Daelemans W. (2008). *Learning the scope of negation in biomedical texts*. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 715–724.

Moreno I., Boldrini E., Moreda P. and Romá-Ferri M. T. (2017). DrugSemantics: a corpus for named entity recognition in spanish summaries of product characteristics. *Journal of Biomedical Informatics* **72**, 8–22.

Moreno-Sandoval A. and Campillos-Llanos L. (2013). Design and annotation of multimedica–A multilingual text corpus of the biomedical domain. *Procedia-Social and Behavioral Sciences* **95**, 33–39.

Mutalik P. G., Deshpande A. and Nadkarni P. M. (2001). Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *Journal of the American Medical Informatics Association* **8**(6), 598–609.

Oronoz M., Gojenola K., Pérez A., de Ilarraza A. D. and Casillas A. (2015). On the creation of a clinical gold standard corpus in spanish: mining adverse drug reactions. *Journal of Biomedical Informatics* **56**, 318–332.

Pabón O. S., Montenegro O., Torrente M., González A. R., Provencio M. and Menasalvas E. (2022). Negation and uncertainty detection in clinical texts written in spanish: a deep learning-based approach. *PeerJ Computer Science* **8**, e913.

Polignano M., de Gemmis M. and Semeraro G. (2021). Comparing transformer-based NER approaches for analysing textual medical diagnoses. In *CLEF (Working Notes)*, pp. 818–833.

Pröllochs N., Feuerriegel S. and Neumann D. (2017). Understanding negations in information processing: Learning from replicating human behavior, arXiv preprint arXiv: 1704.05356.

Qian Z., Li P., Zhu Q., Zhou G., Luo Z. and Luo W. (2016). *Speculation and negation scope detection via convolutional neural networks*. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 815–825.

Ramshaw L. and Marcus M. (1999). Text chunking using transformation-based learning. In *Natural Language Processing Using Very Large Corpora*, pp. 157–176.

Reitan J., Faret J., Gambäck B. and Bungum L. (2015). *Negation scope detection for twitter sentiment analysis*. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 99–108.

Rokach L., Romano R. and Maimon O. (2008). Negation recognition in medical narrative reports. *Information Retrieval* **11**, 499–538.

Sánchez L., Zavala D., Farré-Maduell E., Lima-López S., Miranda-Escalada A. and Krallinger M. (2022). *The SocialDisNER shared task on detection of disease mentions in health-relevant content from social media: methods, evaluation, guidelines and corpora*. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pp. 182–189.

Sandoval A. and Salazar M. (2013). La anotación de la negación e un corpus escrito etiquetado sintáctimanete. Annotation of negation in a written treebank. *Rev. Iberoam. Linguist* **2013**(8), 45–61.

Santiso S., Casillas A., Pérez A., Oronoz M. and Gojenola K. (2014). *Adverse drug event prediction combining shallow analysis and machine learning*. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pp. 85–89.

Segura-Bedmar I., Martínez P., Revert R. and Moreno-Schneider J. (2015). Exploring Spanish health social media for detecting drug effects. *Bmc Medical Informatics and Decision Making* **15**(2), S6.

Solarte-Pabón O., Montenegro O., Blazquez-Herranz A., Saputro H., Rodriguez-González A. and Menasalvas E. (2021). Information extraction from Spanish radiology reports using multilingual BERT. *In Proceedings of CLEF eHealth 2021*, pp. 834–45.

Solarte-Pabón O., Torrente M., Provencio M., Rodríguez-Gonzalez A. and Menasalvas E. (2021). Integrating speculation detection and deep learning to extract lung cancer diagnosis from clinical notes. *Applied Sciences* **11**(2), 865.

https://doi.org/10.1017/nlp.2024.10 Published online by Cambridge University Press

Suárez-Paniagua V., **Dong H. and Casey A.** (2021). *A multi-BERT hybrid system for Named Entity Recognition in Spanish radiology reports*. In *CEUR Workshop Proceedings (Vol. 2936). CEUR Workshop Proceedings.*

Tamayo A., **Angel J. and Gelbukh A.** (2022). Detección del alcance de las negaciones en español usando conditional random fields. *Research in Computing Science* **151**(5), 199–212.

Tamayo A., **Burgos D. A. and Gelbukh A.** (2022). *mbert and simple post-processing: A baseline for disease mention detection in spanish*. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings.*

Tamayo A., **Gelbukh A. and Burgos D. A.** (2022). *Nlp-cic-wfu at socialdisner: Disease mention extraction in spanish tweets using transfer learning and search by propagation*. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pp. 19–22.

Taulé M., **Nofre M.**, **González M. and Martí M.** (2021). Focus of negation: its identification in spanish. *Natural Language Engineering* **27**(2), 131–152.

Vincze V., **Szarvas G.**, **Farkas R.**, **Móra G. and Csirik J.** (2008). The bioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics* **9**(11), 1–9.

White J. (2012). *UWashington: Negation resolution using machine learning methods*. In * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 335–339.

Wu S., **Miller T.**, **Masanz J.**, **Coarr M.**, **Halgrim S.**, **Carrell D. and Clark C.** (2014). Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PloS One* **9**(11), e112774.

Zhu Q., **Li J.**, **Wang H. and Zhou G.** (2010). *A unified framework for scope learning via simplified shallow semantic parsing*. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 714–724.

Zou B., **Zhou G. and Zhu Q.** (2013). *Tree kernel-based negation and speculation scope detection with structured syntactic parse features*. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 968–976.

Cite this article: Tamayo-Herrera AJ, Burgos DA and Gelbukh A (2025). Augmenting a Spanish clinical dataset for transformer-based linking of negations and their out-of-scope references. *Natural Language Processing* **31**, 56–89. https://doi.org/10.1017/nlp.2024.10