Artificial Intelligence In Medicine 145 (2023) 102682

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Negation and speculation processing: A study on cue-scope labelling and assertion classification in Spanish clinical text

Naiara Perez [a,b,1,*], Montse Cuadros [a], German Rigau [b]

[a] SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi Pasealekua 57, Donostia/San Sebastián, 20009, Spain
[b] HiTZ Basque Center for Language Technologies, University of the Basque Country (UPV-EHU), Manuel Lardizabal Ibilbidea 1, Donostia/San Sebastián, 20018, Spain

ARTICLE INFO

Keywords:
Deep learning
Natural Language Processing
Biomedical language processing
Negation
Speculation

ABSTRACT

Natural Language Processing (NLP) based on new deep learning technology is contributing to the emergence of powerful solutions that help healthcare providers and researchers discover valuable patterns within insurmountable volumes of health records and scientific literature. Fundamental to the success of such solutions is the processing of negation and speculation. The article addresses this problem with state-of-the-art deep learning approaches from two perspectives: cue and scope labelling, and assertion classification. In light of the real struggle to access clinical annotated data, the study *(a)* proposes a methodology to automatically convert cue-scope annotations to assertion annotations; and *(b)* includes a range of scenarios with varying amounts of training data and adversarial test examples. The results expose the clear advantage of Transformer-based models in this regard, managing to overpass a series of baselines and the related work in the public corpus NUBes of clinical Spanish text.

1. Introduction

Natural Language Processing (NLP) based on new deep learning technology is contributing to the emergence of powerful information extraction and retrieval solutions for healthcare providers and researchers [1], for instance, to discover valuable patterns within the ever increasing volumes of health records and scientific medical literature. Fundamental to the success of such solutions in the medical domain is the processing of negation and speculation. This work focuses on the automatic detection of negation and speculation in health records.

Negation is the universal linguistic phenomenon that reverses the polarity of statements or clauses, most typically by the usage of words like “no” or “not”. *Speculation* has to do with modality. In this work and the related studies, it is an umbrella term that refers broadly to linguistic phenomena related to hedging, evidentiality, uncertainty, and factuality [2]. To put it simply, we construe speculation as explicit language that signals a speaker is unsure whether a statement is true or lacks evidence to commit fully to it.

Several clinical corpora descriptions [3–8] report incidences of negation and speculative language in, respectively, 10%–35% and 5%–15% of the analysed sentences. That is, up to half the sentences of clinical narrative could potentially contain these type of linguistic constructs, which transform entirely the meaning of the texts they appear in. Properly detecting and handling them is thus a crucial feature of any NLP solution aimed at assisting the clinical practice through the exploitation of clinical narrative.

The NLP community has proposed multiple models to represent the problem of negation and speculation detection:

On the one hand, there is the task of *detecting cues and scopes*, the constituent parts of negation and speculation, as pictured in Fig. 1. *Cues* (also called *markers* or *triggers*) are words or phrases that express negation or speculation. *Scopes* are the clauses affected by a cue, that is, whose propositional values are somehow modified. The detection of cues and/or scopes is usually addressed as a sequence labelling problem. Some works focus exclusively on finding the scopes of given pre-annotated cues; this task is known as negation and/or speculation *scope resolution*.

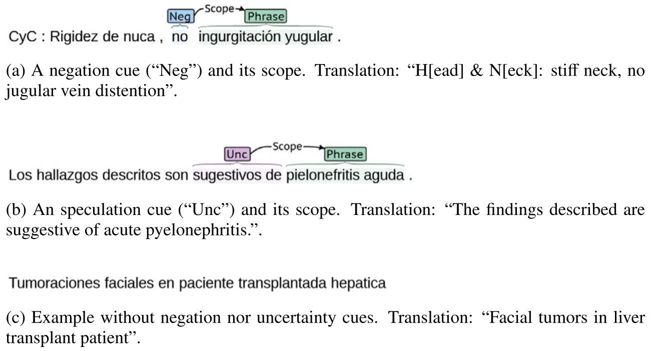
The second common way of modelling negation and uncertainty detection in the biomedical field is as a text classification task known as *assertion classification*. In this case, the text to analyse is pre-annotated with medical entities, whose *assertion category* – present, absent, or possible – needs to be determined. The sentences of Fig. 1 are depicted in Fig. 2 framed as entity assertion annotations.

[*] Corresponding author at: HiTZ Basque Center for Language Technologies, University of the Basque Country (UPV-EHU), Manuel Lardizabal Ibilbidea 1, Donostia/San Sebastián, 20018, Spain.

E-mail address: naiara.perez@ehu.eus (N. Perez).

[1] The work was carried out in Vicomtech.

https://doi.org/10.1016/j.artmed.2023.102682
Received 10 June 2022; Received in revised form 25 August 2023; Accepted 6 October 2023
Available online 10 October 2023
0933-3657/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

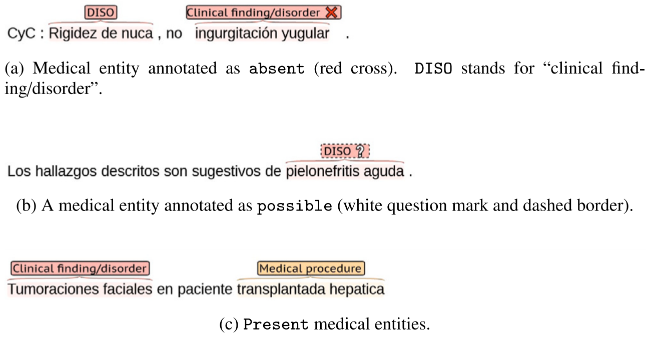

Neg
Scope
Phrase
CyC : Rigidez de nuca , no ingurgitación yugular .
Unc
Scope
Phrase
Los hallazgos descritos son sugestivos de pielonefritis aguda .
Tumoraciones faciales en paciente transplantada hepatica

(a) A negation cue (“Neg”) and its scope. Translation: “H[ead] & N[eck]: stiff neck, no jugular vein distention”.

(b) An speculation cue (“Unc”) and its scope. Translation: “The findings described are suggestive of acute pyelonephritis.”.

(c) Example without negation nor uncertainty cues. Translation: “Facial tumors in liver transplant patient”.

Fig. 1. Annotations of negation and speculation cues and scopes.


DISO
Clinical finding/disorder ✗
CyC : Rigidez de nuca , no ingurgitación yugular .
DISO ?
Los hallazgos descritos son sugestivos de pielonefritis aguda .
Clinical finding/disorder
Medical procedure
Tumoraciones faciales en paciente transplantada hepatica

(a) Medical entity annotated as `absent` (red cross). `DISO` stands for “clinical finding/disorder”.

(b) A medical entity annotated as `possible` (white question mark and dashed border).

(c) `Present` medical entities.

Fig. 2. Annotations of medical entities and their assertion category (see translations in Fig. 1).

While the automatic processing of negation is a well-studied problem, the detection of speculation has received much less attention in general. Furthermore, the vast majority of studies focus on English text, as is usually the case [9,10]. Here, we address both the problems of detecting negation and speculation. The study focuses in Spanish text of the health domain through the public corpus NUBes [8]. Spanish is the official language of 21 countries and the second most spoken language in the world by number of native speakers (around 493 million), surpassing English [11]. Thus, accurately processing the large volumes of digital health records in Spanish is of the most importance. The specific contributions of our work are the following:

- Addressing the task of assertion classification (in addition to cue and scope detection), which has not been tackled in Spanish clinical text since the surge of deep learning in NLP. As a novelty, we tackle both tasks through the same corpus. To that end, we propose a series of steps to convert a corpus such as NUBes, annotated with cues and scopes, to a corpus suitable for assertion classification purposes.
- Exploiting a diverse set of Transformer models and Flair models in both tasks, the performance of which we compare to baseline results and related work. We manage to improve the state of the art in the detection of cues and scopes in the NUBes corpus, most markedly in regards with speculation, the most challenging category.
- Analysing the performance of said models in a range of scenarios of varying difficulty:
    a) In addition to the overall performance a given model may yield, being able to achieve competitive results with as little data as possible is a most desirable trait, given that clinical data is notably hard to obtain. For this reason, we analyse the performance of the models with decreasing amounts of training data, from thousands of examples down to a few dozen.
    b) It has been widely reported that a few negation markers (e.g., “no” and “sin”) are responsible for most of the negation instances in Spanish free text [5,8,12,13]. While previous studies on negation and uncertainty detection report overall acceptable results in multiple scenarios and datasets, it has not been studied how well predictive models perform specifically on the less frequent surface forms of negation, which are equally important in real usage scenarios.
- Providing a comprehensive survey of the related work and exposing its lack of comparability. As is well-known, clinical NLP suffers from a generalised impossibility to make data public due to privacy issues, which is in itself the major impediment for comparability. On top of that, there is a lack of consensus in previous related work regarding the evaluation metrics. To compare ourselves to others, and to facilitate future comparisons, we report our results following 3 distinct evaluation methodologies.

2. Background

Several survey articles report, particularly for English, on the research of automatic negation and speculation processing [14–16]. In the last years, the processing of negation in Spanish text has also gained attention encouraged by the NEGES (Negation in Spanish) workshops [17,18] and the publications of several freely available corpora: IULA-SCRC [7], SFU Review$_{SP}$-NEG [19], NUBes [8], NewsCom [20], and T-MexNeg [21], These corpora differ in text genre and domain, and conform to divergent guidelines for string-level annotations of negation cues, scopes and events. IULA-SCRC [7] and SFU Review$_{SP}$-NEG [19] are the most used in the literature, the former being from the clinical domain and the latter from the product-reviews domain. NewsCom [20] is a corpus of online comments posted in response to news articles, and T-MexNeg [21] is made of tweets written in Mexican Spanish. NUBes [8], curated from medical reports, is the only one that also considers speculation along with negation.

Several approaches have been applied to the automatic processing of negation and speculation in Spanish, including hand-crafted heuristics, shallow machine learning and, more recently, deep learning. Table 1 offers a summary of this work, which we present below; of note, the table also exposes how fragmented this research field is, the only comparable results being those pertaining to the NEGES workshops [17,18] or having been authored by the same researchers.

The earliest related studies [22–26] consist of different adaptations and/or extensions of NegEx [27] to the Spanish language. NegEx is an algorithm originally based on English lexicons that categorises pre-annotated medical entities as present or absent given the contexts the entities occur in. These Spanish adaptations obtain F1-scores 0.64 to 0.78.

Koza et al. [28] worked on the recognition of negated medical findings in radiological reports by means of rules based on morphosyntactic and semantic information. They report an F1-score of 0.98 on an evaluation against their own private corpus, but acknowledge that the test data set lacks variability in the negation structures it includes.

The task of recognising negated findings has also been undertaken by Santiso et al. [29,30], but with machine learning techniques and modelling the problem as a sequence labelling task. They first assess Conditional Random Fields (CRFs) [31] over symbolic features and features derived from word embeddings, achieving 0.82 and 0.75 span-level F1-score (partial match) in IULA-SCRC [7] and their private corpus IxaMed-GS [32], respectively. Next, they implement a Recurrent Neural Network (RNN) featuring character embeddings, bidirectional Long Short-Term Memory (bi-LSTM) layers and a CRFs classifier, surpassing their previous results on IxaMed-GS.

Table 1
Literature review on negation and uncertainty detection in Spanish text. *SEM 2012 F1 is the evaluation metric proposed by Morante and Blanco [39] for the *SEM 2012 shared task on resolving the scope and focus and negation. ZS stands for zero-shot performance. Notice that scores are only comparable if they result from the same evaluation corpus, task and metric. An extensive discussion of the different evaluation metrics can be consulted in Sineva et al. [40].

Evaluation corpus	Ref	Task	System	Metric	Score
SFU Review$_{SP}$-NEG [19]	[33]	NEG cue detection	CRF	*SEM 2012 F1	0.86
	[34]	NEG cue detection	bi-LSTM	*SEM 2012 F1	0.68
	[35]	NEG cue detection	bi-LSTM	*SEM 2012 F1	0.83
	[36]	NEG cue detection	CRF	*SEM 2012 F1	0.84
	[37]	NEG cue detection	CRF	*SEM 2012 F1	0.81
	[38]	NEG cue detection	bi-GRU	*SEM 2012 F1	0.23
	[41]	NEG cue detection	CRF	*SEM 2012 F1	0.87
	"	NEG scope resolution	CRF	*SEM 2012 F1	0.81
	[42]	NEG scope resolution	Transformer (ZS)	token F1	0.78
	[43]	NEG scope resolution	Transformer (ZS)	token F1	0.79
	[44]	NEG cue and scope detection	Transformer	*SEM 2012 F1	0.88
IULA-SCRC [7]	[45]	NEG scope resolution	Transformer (ZS)	*SEM 2012 F1	0.94
	[26]	NEG cue and scope detection	Rules	sentence F1	0.92
	[44]	NEG cue and scope detection	bi-LSTM + CRF	CoNLL-2010 scope F1	0.85
	[29]	negated entity detection	CRF	inexact span F1	0.82
	[46]	NEG scope detection	Transformer	BIO-weighted token F1	0.88
NUBes [8]	[45]	NEG scope resolution	Transformer (ZS)	*SEM 2012 F1	0.90
	[8]	NEG cue detection	bi-LSTM + CRF	token F1	0.96
	"	UNC cue detection	bi-LSTM + CRF	token F1	0.85
	"	NEG scope detection	bi-LSTM + CRF	token F1	0.91
	"	UNC scope detection	bi-LSTM + CRF	token F1	0.79
	[46]	NEG cue detection	Transformer	BIO-weighted token F1	0.95
	"	UNC cue detection	Transformer	BIO-weighted token F1	0.84
	"	NEG scope detection	Transformer	BIO-weighted token F1	0.88
	"	UNC scope detection	Transformer	BIO-weighted token F1	0.72
Private corpora	[22]	assertion classification	Rules	F1	0.74
	[23]	assertion classification	Rules	F1	0.67
	[28]	negated entity detection	Rules	sentence F1	0.98
	[24]	negated entity detection	CRF + Rules	inexact span F1	0.74
	[29]	negated entity detection	CRF	inexact span F1	0.75
	[30]	negated entity detection	bi-LSTM + CRF	inexact span F1	0.82
	[46]	NEG cue detection	Transformer (ZS)	BIO-weighted token F1	0.90
	"	UNC cue detection	Transformer (ZS)	BIO-weighted token F1	0.81
	"	NEG scope detection	Transformer (ZS)	BIO-weighted token F1	0.84
	"	UNC scope detection	Transformer (ZS)	BIO-weighted token F1	0.74

Systems based on CRFs and bi-LSTMs were also the most popular among the participants of the shared task about negation cue detection in the NEGES workshops [33–38]. The corpus provided in both workshop editions to train and test the competing systems was SFU Review$_{SP}$-NEG [19]. The best overall results (0.86 span-level F1-score) were obtained by Loharja et al. [33] with a CRFs classifier over lexical and morphological features.

The organisers of NEGES implemented another CRFs classifier improving the state of the art on negation cue detection in SFU Review$_{SP}$-NEG with an F1-score of 0.87 [41]. This is also the first work in the literature that tackles the problem of negation scope resolution along with cue detection in Spanish text. Specifically, they follow a 2-stage setup with two separate classifiers, where the first detects cues, whose scopes are determined by the second. The classifier of scopes yields F1-scores of 0.81 and 0.73 with gold and predicted cues as input, respectively.

In view of the across-the-board success of the Transformer architecture [47] and the availability of pre-trained neural language models steadily increasing in number and size, the focus of works about negation detection has recently shifted towards studying how these large pre-trained language models behave and what advantages they offer.

Rivera Zavala and Martínez [44] compare a RNN-based classifier and a Transformer-based classifier in the task of negation cue detection and scope resolution in the corpora IULA and SFU Review$_{SP}$-NEG. The RNN classifier combines character, word and sense embeddings as input to a bi-LSTM network, whose output is fed to a CRFs classifier. The BERT-based system follows the conventional setup of a pre-trained language model (Multilingual BERT o mBERT[2]) with a softmax output layer. Both systems tackle the problem of cue and scope detection jointly. They achieve 0.81 and 0.85 token-level F1-score with BERT and the RNN, respectively, in the IULA-SCRC corpus. In SFU Review$_{SP}$-NEG, the results are 0.92 and 0.88.

Shaitarova et al. [42], Shaitarova and Rinaldi [43] explore the transferability of negation scope resolution models between the languages English, French, Spanish and Russian. Their work is built on NegBERT [48], a system originally built for English that performs negation cue detection and scope resolution in a 2-stage fashion using BERT. These works adapt NegBERT to the cross-lingual setting by replacing BERT with mBERT and XLM-RobERTa [49]. They achieve token-level F1-scores ∼0.78 when zero-shot testing English and French models on the SFU Review$_{SP}$-NEG corpus, with XLM-RobERTa outperforming mBERT by a narrow margin.

Hartmann and Søgaard [45] also study zero-shot cross-lingual transfer approaches for negation scope resolution. Specifically, they explore how to best exploit disparate available datasets (in their work, multiple datasets in English) to overcome the lack of training data on the target languages (here, Spanish). They propose the application of a Multi-Task Deep Neural Network (MT-DNN) [50], where each dataset available for training is treated as an independent task. This approach is compared to the simple concatenation of the training datasets, which they find works slightly better overall when evaluated in IULA-SCRC [7] and NUBes [8], among others. They report *SEM 2012 scope token F1-scores [39] of 0.94 and 0.90 in these datasets, respectively.

Notably, the processing of speculation is yet to be thoroughly addressed in Spanish text (clinical or otherwise). This task is considerably more challenging than the detection of negation cues and scopes [3,8,46,51], due the context-dependent subtlety, lexical variety, and gradation of cues for uncertainty, which are also less frequent than

[2] https://github.com/google-research/bert/blob/master/multilingual.md

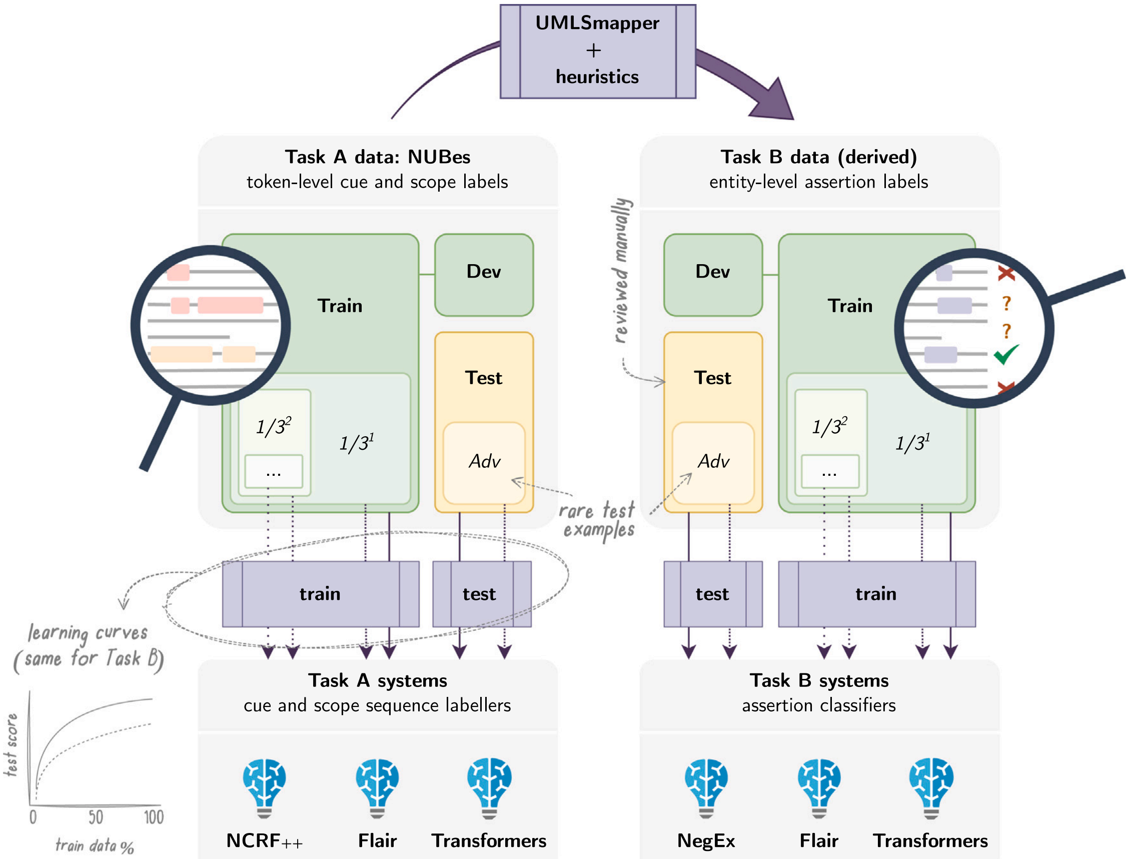


Fig. 3. Experimentation flowchart, starting from the NUBes corpus, on cue and scope labelling (Task A) and assertion classification (Task B).

negation. Lima-López et al. [8] report the first exploratory experiments with the NUBes corpus using the biLSTM + CRF architecture over a rich set of morphosyntatic and lexical features. This work has recently been extended to incorporate the first published experiments with a Transformer-based model on the NUBes corpus [46], achieving similar results to Lima-López et al. [8].

In this work, we test on NUBes the latest and more competitive neural language models for Spanish. We compare them to multiple baselines and the related work [8,45,46], which we manage to overpass in the setting of supervised negation and uncertainty cue and scope detection. In order to be able to compare ourselves with these works, we report our results in the various corresponding metrics. Most interestingly, we use the NUBes corpus to study the task of assertion classification as well, which has not been tackled in Spanish clinical text since the surge of deep learning in NLP.

3. Materials and methods

This article describes two sets of parallel experiments on the detection of negation and speculation:

- **Task A:** cue and scope detection (introduced in Fig. 1), framed as a sequence labelling problem.
- **Task B:** assertion classification (Fig. 2), framed as a text classification problem.

In each case, a series of systems are fine-tuned, trained or adapted with the training set of the NUBes corpus [8], to be next evaluated against the testing set of the same corpus. That is, we focus on supervised techniques and report in-domain results.

In addition, we assess the data requirements of the models by observing their *learning curves*, that is, by training or fine-tuning them on decreasing amounts of data. This is particularly relevant in the medical domain, where accessing clinical narrative, annotated or otherwise, is generally problematic.

Furthermore, the above mentioned evaluations include two testing sets, henceforth referred to as Full and Adv (from “adversarial”). As is explained in detail below, Adv contains exclusively the less common surface forms of negation and speculation, while Full is a regular random sample of NUBes.

The accompanying flowchart (Fig. 3) visually outlines the various stages of this experimentation process. In the upcoming subsections, we will delve into the specifics of the training and test data, elaborate on the evaluated systems (with architectures, implementation, and training specifics), and finally, outline the performance metrics used to measure the models’ performance.

3.1. Data

The experiments are conducted with the NUBes corpus [8]. It consists of a collection of sentences extracted from anonymous Spanish clinical records and manually annotated with negation and uncertainty cues and their scopes. That is, originally, this corpus is meant to be used for Task A. The next section describes the corpus as is, while Section 3.1.2 explains how we generated a new corpus from NUBes for the assertion classification experiments (Task B).

3.1.1. Data of Task A: cue and scope detection

For this set of experiments we use the original train, development and testing splits of the NUBes corpus,[3] which already come tokenised and tagged with 4 types of entities:

[3] https://github.com/Vicomtech/NUBes-negation-uncertainty-biomedical-corpus

Table 2
Size of the corpus for Task A – cue and scope detection.

	Train	Dev	Test	
			Full	Adv
Sentences	13,802	1,840	2,762	1,838
with negation	5,265	694	1,041	240
with uncertainty	1,272	162	249	206
with both	364	64	91	11
Negation cue (NC)	6,976	919	1,423	265
Negation scope (NS)	6,379	847	1,322	233
Uncertainty cue (UC)	1,866	263	400	251
Uncertainty scope (US)	1,886	260	400	249
Total	17,107	2,289	3,545	998

Table 3
Cues with relative frequency > 2% on the train set of Task A.

	Type	#	%	C%
"no"	negation	3,046	34.35	34.35
"sin"	negation	1,820	20.53	54.88
"probable"	uncertainty	264	2.98	57.86
"afebril"	negation	190	2.14	60.00
"asintomático"	negation	187	2.11	62.11

- NC: negation cue,
- NS: negation scope,
- UC: uncertainty cue, and
- US: uncertainty scope.

The NUBes annotations follow the BIO scheme [52], in which B marks the beginning of a span, while the subsequent tokens of the span receive the tag I (from "in") and tokens that do not belong to any span are marked with O ("out"). This setting yields a total of 9 possible tags per token. The sentences of Fig. 1 would be encoded as follows with the presented tagset:

(3.1). *From Fig. 1(a):*

CyC	O
:	O
Rigidez	O
de	O
nuca	O
,	O
no	B-NC
ingurgitación	B-NS
yugular	I-NS
.	O

(3.2). *From Fig. 1(b):*

Los	O
hallazgos	O
descritos	O
son	O
sugestivos	B-UC
de	I-UC
pielonefritis	B-US
aguda	I-US
.	O

The total size of each data split can be consulted in Table 2. To compute the train curves, we create increasingly smaller training data subsets by randomly extracting 1/3 of the examples for 5 iterations, for a total of 6 decremental training datasets (see Appendix A).

To create the difficult or adversarial test data set, Adv, we remove from the original test data set (Full) the examples that contain frequent negation or speculation cues, being frequent any cue with relative frequency in the training set higher than 2%, which together constitute 62.11% of the cues (see Table 3). That is, Adv is a subset of the regular test set.

As can be seen, negation instances are more than thrice more likely to occur than speculation in this corpus; furthermore, speculation cues are lexically more variable, as evidenced by the smaller drop from the regular to the difficult test set.

3.1.2. Data of Task B: assertion classification

In the task of assertion classification, each instance consists of the medical entity to be classified presented in context. The categories of the task are the following:

- absent or abs: negated medical entity,
- possible or pos: uncertain medical entity, and
- present or pre: positive medical entity.

From the examples in Fig. 2, we would get the following instances (one per medical entity):

(3.3). *CyC: <e>**Rigidez de nuca**</e>, no ingurgitación yugular.* pre

(3.4). *CyC: Rigidez de nuca, no <e>**ingurgitación yugular**</e>.* abs

(3.5). *Los hallazgos [...] sugestivos de <e>**pielonefritis aguda**</e>.* pos

(3.6). *<e>**Tumoraciones faciales**</e> en paciente [...]* .... pre

(3.7). *[...] en paciente <e>**transplantada hepatica**</e>* ...... pre

At the moment of executing the experiments described here, there is no publicly available dataset in Spanish annotated with medical entities and their assertion category. Thus, in order to conduct this experiment, we automatically construct a new corpus from NUBes, with the help of the original cue, scope and entity annotations.[4] The transformation process is as follows:

First, we automatically annotate the entire corpus with medical entities. To that end we exploit UMLSmapper [53,54], a tool for annotating medical entities in Spanish texts and linking them to the UMLS Metathesaurus [55]. Specifically, we annotate mentions of the following types of entities: clinical findings and disorders, procedures, chemicals and drugs, physiological phenomena, and some living beings (namely viruses, bacteria, and fungi).[5]

Then, we automatically assign the categories absent, possible or present to each annotated entity depending on whether they occur within the scope of a negation cue, an speculation cue or neither, respectively.

To be specific, however, not all the entities that fall within the scope of a negation or speculation cue are directly affected by it. Consider the sentence in Fig. 4. While "secuela quirurgica" is a clinical finding under the scope of an uncertainty cue, the speculation is rather about the facial paresis than the surgical sequelae or the relation of the former to the latter. NUBes comes with manual annotations of entities, but only of those most prominently affected by the corresponding cue. Based on this information, we remove the entities that fall within the scope of a cue but that do not overlap with a manually annotated entity in the cases there is one. This way, we avoid incorrectly annotating as negated or uncertain entities such as "secuela quirurgica" in Fig. 4.

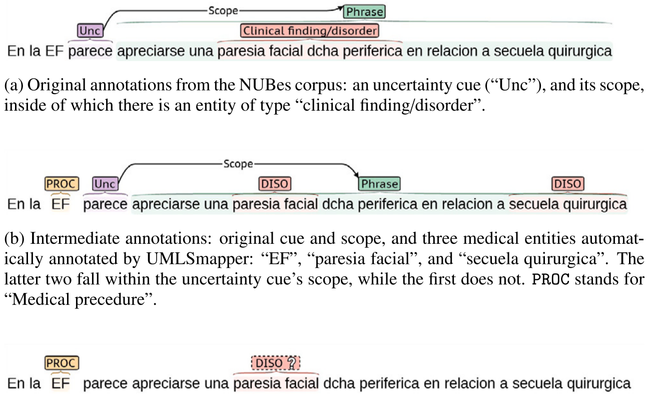
Even then, we have manually revised the testing portion of the dataset, which allows us, on the one hand, to measure the validity of the proposed data conversion and, on the other hand, to ensure the reliability of the reported results and conclusions drawn therefrom. The manual revision led to correcting the assertion category of 38 instances and removing 7 instances out of the 2,474 revised examples.

Finally, each annotated entity must be converted to the text classification format presented earlier (see Example 3.3 to 3.7). The two entity annotations in Fig. 4(c) would yield the following instances:

(3.8). *En la <e>**EF**</e> parece apreciarse una paresia facial [...]* pre

[4] The resulting dataset can be found at the NUBes repository.

[5] The classification of types is given by the UMLS semantic groups [56]. Notice that we do not care about the correctness of the UMLS links established by UMLSmapper nor of the entity types assigned thereof, which we simply use to filter the annotations. The task the classification models need to learn is to establish a relation between the entity and the context it occurs in, in order to emit a prediction regarding whether the entity is present, absent, or possible. The type of the entity (disorder, drug, and so on) is irrelevant to the task, even more so its link to the UMLS Metathesaurus.


Scope
Phrase
Unc
Clinical finding/disorder
En la EF parece apreciarse una paresia facial dcha periferica en relacion a secuela quirurgica
Scope
PROC
Unc
DISO
Phrase
DISO
En la EF parece apreciarse una paresia facial dcha periferica en relacion a secuela quirurgica
PROC
DISO ?
En la EF parece apreciarse una paresia facial dcha periferica en relacion a secuela quirurgica

(a) Original annotations from the NUBes corpus: an uncertainty cue (“Unc”), and its scope, inside of which there is an entity of type “clinical finding/disorder”.

(b) Intermediate annotations: original cue and scope, and three medical entities automatically annotated by UMLSmapper: “EF”, “paresia facial”, and “secuela quirurgica”. The latter two fall within the uncertainty cue’s scope, while the first does not. `PROC` stands for “Medical precedure”.

(c) Final processed annotations of medical entities and their assertion status: “EF” is `present` while “paresia facial” is `possible`.

Fig. 4. Steps to transform a NUBes instance for the entity classification corpus. Translation: “In the P[hysical] E[xamination], a peripheral right facial paresis is seemingly noticed in relation to surgical sequelae”..

Table 4
Size of the corpus for Task B – assertion classification.

	Train	Dev	Test		
			Full	Adv	Man
`negated`	2,399	331	460	95	973
`uncertain`	1,001	140	197	125	327
`present`	8,708	1,188	1,810	1,287	0
out-of-scope	3,912	534	818	295	0
from assertion	4,796	654	992	992	0
Total	12,108	1,659	2,467	1,507	1,300

(3.9). *En la EF parece apreciarse una <e>**paresia facial**</e> [...]* `pos`

In this experiment set, we also work with the original training, development and test splits of the NUBes corpus, as in [8]. The resulting dataset is described quantitatively in Table 4. We applied the same methodology as for Task A to generate incremental training subsets (1/1 through $1/3^5$) and the more difficult testing set, Adv, as explained in the previous Section 3.1.1.

In addition, Task B also exploits the original entity annotations of the NUBes corpus, that is, the manual (Man) annotations of entities. This test set is simply added for the sake of completeness, although it does not include `present` annotations—which is why the corpus had to be automatically re-annotated.

Of note, Table 4 breaks down examples of `present` findings into two categories: *out-of-scope* and *from assertion*. The former are examples of entities mentioned in the context of a negation or speculation cue, but that are not affected by it (e.g., “EF” in Fig. 4(c)); the latter are examples generated from sentences without negation nor uncertainty. Without out-of-scope examples, the models would simply learn to detect the presence or absence of negation and uncertainty cues, regardless of whether they affect or not the target entity to be classified.

3.2. Systems

Cue and scope detection (Task A) has been framed as a sequence labelling problem. The trained sequence labellers learn to detect jointly the 4 span types as a single task, emitting for each input token one of the 9 defined labels (see Section 3.1.1). On the other hand, assertion classification (Task B) is a text classification task, where each medical entity whose assertion status needs to be predicted is presented in context one by one to the systems (see Examples 3.8 and 3.9). Each experiment set involves a baseline system, a Flair-based system and several Transformer-based systems, which we present below. Implementation details are given at the end of the section.

3.2.1. NCRF++

The baseline for cue and scope detection (Task A) was set by Lima-López et al. [8] with the NCRF++ [57] sequence tagger. In few words, the system consists of a Convolutional Neural Network (CNN) layer for character sequence representations, which are concatenated to word and feature embeddings, then fed to a bi-LSTM layer and an output CRF layer. The character, word, and feature embeddings are initialised randomly and trained on the given corpus. Here, we report the results of the best variant produced by Lima-López et al. [8], which operates on a set of lexical and morpho-syntactic features automatically extracted from the input text.

3.2.2. NegEx

As is customary in assertion classification, the NegEx [27] system serves as a baseline in our experiments of *Task B*. NegEx is a rule-based system that leverages hand-crafted lexicons in order to determine the assertion categories of the given entities.

The lexicons define 4 types of words or expressions: conjunctions, pseudo-negation cues, negation cues and speculation cues. The first two are used to find the boundaries of scopes and to discard false cues, respectively. Negation and speculation cues are further divided into two groups each, depending on whether they precede (PRE) or follow (POST) their scopes.

Although NegEx has been adapted to Spanish in several occasions (see Section 2), only one adaptation is publicly available [25]. Unfortunately, it does not consider speculation. Thus, we use the original NegEx Python implementation[6] with cues automatically extracted from our training data sets. The categories of the cues (PRE or POST) are automatically determined by choosing the most frequent position in the corpus.[7] The conjunction and pseudo-negation lexicons have been taken from Santamaría [25] as is.

3.2.3. Flair

Flair is a NLP Python framework [58] that features a specific type of character-based contextualised word embeddings of the same name [59]. Here we train Flair’s sequence tagger for Task A and text classifier for Task B following the official documentation recipes,[8] which we explain briefly below.

Both architectures have in common the input embedding mechanism: we use Flair’s pre-trained embeddings for Spanish (`es-forward` and `es-backward`) in combination with the fastText embeddings [60] Biomedical Word Embeddings for Spanish[9] (BWES) [61]. Both are updated during training.

Then, the sequence tagger passes the embeddings to a bi-LSTM layer and an output CRF layer. In short, the main differences of this system with respect to the baseline sequence labeller are that *(a)* it uses pre-trained contextual character embeddings instead of static embeddings trained from scratch, and *(b)* it starts off with language and domain knowledge.

In the case of the text classifier, the computed embeddings are fed into a Gated Recurrent Unit (GRU) layer to produce a document level representation, which is then passed to a classification layer to make the assertion category prediction. A simplified diagram of the Flair-based text classifier for Task B is shown in Fig. 5.

[6] https://github.com/chapmanbe/negex

[7] PRE and POST lexicons are also available at the NUBes repository.

[8] https://github.com/flairNLP/flair/tree/master/resources/docs

[9] https://github.com/TeMU-BSC/Embeddings (v2.0, Skip-gram, SciELO+Wiki, uncased)

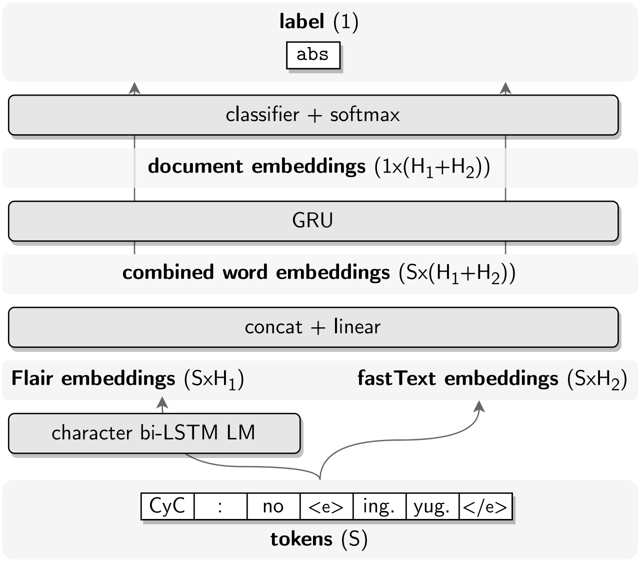


Fig. 5. Diagram of the Flair-based assertion classifier. S (sequence length); H_1 = 128 or 256 (Flair embedding size); H_2 = 300 (fastText embedding size).

3.2.4. Transformer

The bulk of the experimentation involves Transformer [47] models. It includes a diverse set of BERT- [62] and RoBERTa-like [63] pre-trained language models, both monolingual and multilingual, as well as general-purpose and domain-specific. Further, each of the selected pre-trained language models serves to train a sequence tagger and a text classifier for Task A and Task B, respectively. We used HuggingFace's Trainer implementations,[10] which we explain briefly below.

The architecture of the sequence tagger follows the standard layer stack: the BERT encoder is followed by a dropout layer and one classification head consisting of a linear transformation layer, which emits the logits per token for the 9 possible categories. The models are trained on the cross-entropy loss of the classification head over the first subword of each input token. For inference, the label with the maximum probability is chosen for each token after applying the softmax function to the logits. Subwords in suffix positions are ignored, that is, the label for each token is assigned from the prediction for the first subword. Fig. 6 illustrates this architecture with a simplified diagram.

As for the text classification architecture, it differs from the sequence tagger in that the classifier head is fed the pooled output of the encoder. The pooled output is computed over the special token at the beginning of each sequence (i.e., BERT's `[CLS]` and RoBERTa's `<s>`) by passing its embeddings to a dense linear layer and a tanh activation function. The result is then fed to a dropout layer and the final dense linear layer, which outputs the logits for the 3 assertion categories. For this task, we added the special tokens `<e>` and `</e>`, which mark the start and end of the medical entity, to the vocabularies of the pre-trained models.

The full list of tested pre-trained models can be consulted in Table 5. Table 6 describes those models in terms of their vocabulary overlap with NUBes. For comparison purposes, the same table reports the vocabulary overlap with SFU Review$_{SP}$-NEG [19], a corpus of product reviews in Spanish. As can be seen, the greatest vocabulary coverage, provided by SpanBERTa, is 28.47%. That is, 28.47% of the set of words occurring in NUBes have their own embedding. When weighted by word frequency, the coverage rises to 69.67% of the corpus (ignoring stopwords). The worst model in this regard is, unsurprisingly, SciBERT [66]—a monolingual English model—, with just 6.02% vocabulary overlap with NUBes.

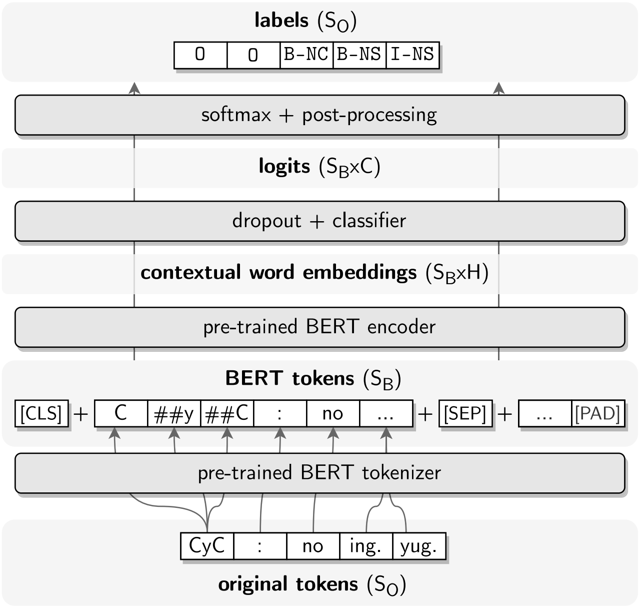


Fig. 6. Diagram of the BERT-based cue and scope tagger. S_O (original sequence length); S_B = 220 (sequence length after BERT tokenisation and padding); H = 768 (BERT embedding size); C = 9 (number of output labels).

3.2.5. Implementation and training setup

We have optimised some hyperparameters of the Transformer variants and Flair in each task and training data subset with 25 trials each. The Transformer models have been implemented with HugginFace's `transformers` Python library [69] and optimised using Ray's `tune` Python library [70]. In the case of Flair, the Python library comes with a wrapper[11] of Hyperopt [71] for hyperparameter optimisation. In each case, the trial with the best F1-score on the development data set has been used to compute the results on the testing data sets. The hyperparameter search spaces are reported in Appendix B.

As for the baseline systems, in the case of NegEx, we compute the learning curve by extracting the negation and speculation cues only from the corresponding training data subset at each point. The NCRF++ tagger is the same as that described by Lima-López et al. [8]. We also include its hyperparemeter setup in Appendix B for convenience.

3.3. Evaluation

Both tasks are evaluated in terms of F1-score ($F1$), the harmonic mean of Precision (P) and Recall (R):

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{1}$$

where the true positives (TP), false positives (FP) and false negatives (FN) are defined differently for each task, as explained below. The three metrics reach their best value at 1. Intuitively, Recall measures how many gold instances have been correctly predicted, while Precision measures how correct the predictions made are.

In the case of negation and uncertainty cue and scope detection (Task A), we report *strict span-level metrics* as computed by the publicly available, open-source Python library `seqeval`.[12] The token-level

[10] https://github.com/huggingface/transformers/blob/main/examples/pytorch

[11] https://github.com/flairNLP/flair/blob/master/resources/docs/TUTORIAL_8_MODEL_OPTIMIZATION.md

[12] https://github.com/chakki-works/seqeval.

Table 5
Pre-trained language models tested in the experimentation.

		Language	Domain	Corpus	Params	Vocabulary
BERTs	BETO$_{Base}$ Cased [64]	es	generic	Spanish Corpora[a]	110M	31,002
	mBERT$_{Base}$ Cased[b]	multi (104)	generic	Wikipedia	178M	119,547
	IXAmBERT$_{Base}$ Cased [65]	es, en, eu	generic	Wikipedia	178M	119,101
	SciBERT$_{scivocab}$ Cased [66]	en	scientific	Semantic Scholar	110M	31,116
RoBERTas	SpanBERTa$_{Base}$ Cased[c]	es	generic	OSCAR [67]	125M	50,265
	MarIA RoBERTa$_{Base}$ BNE [68]	es	generic	BNE selective crawls[d]	125M	50,262
	XLM-RoBERTa$_{Base}$ [49]	multi (100)	generic	Common Crawl	278M	250,002

[a] https://github.com/josecannete/spanish-corpora.

[b] https://github.com/google-research/bert/blob/master/multilingual.md.

[c] https://github.com/chriskhanhtran/spanish-bert.

[d] http://www.bne.es/en/Colecciones/ArchivoWeb/Subcolecciones/selectivas.html.

Table 6
Vocabulary coverage by the pre-trained language models. UNK is the percentage of unique words in the corpus for which the tokenizer yielded the special token `[UNK]` (or analogous). SHA is the percentage of unique words in the corpus that is covered by the vocabulary. WSH is the percentage of all the words in the corpus (i.e., frequency weighted unique words) that is covered by the vocabulary, after removing stopwords. The models are shown by weighted coverage in the NUBes corpus in descending order.

	NUBes			SFU Review$_{SP}$-NEG		
	UNK	SHA	WSH	UNK	SHA	WSH
SpanBERTa	0.00	28.47	69.67	0.00	55.44	86.47
IXAmBERT	0.73	25.63	66.84	0.55	49.10	79.41
BETO	0.78	21.72	62.25	0.30	41.05	77.12
MarIA	0.00	26.17	51.71	0.00	51.42	63.13
mBERT	0.00	12.97	50.56	0.04	25.32	63.75
XLM-R	0.00	14.40	38.68	0.00	26.00	49.65
SciBERT	0.24	6.02	29.93	0.11	7.99	33.12

predictions are first converted to span-level predictions, that is, the BIO tags are interpreted to obtain predictions consisting of a span boundaries (offset and end) and the predicted category for the span. Then, TP, FP and FN are computed per category $c \in$ {`NC` `NS`, `UC`, `US`} as follows:

- TP: number of predicted spans of category c that match exactly in boundaries with a gold span of category c.
- FP: number of predicted spans of category c that do not match exactly in boundaries with any gold span or that match with a gold span of a category other than c.
- FN: number of gold spans of category c that do not match exactly in boundaries with any predicted span or that match with a predicted span of a category other than c.

This is the strictest evaluation methodology possible for this task. In order to be able to compare the results with the related work, we also report the performances of the trained sequence labelling systems following two additional evaluation methodologies, namely *SEM 2012 scores [39] and BIO-weighted token-level scores [46]. We refer the reader to the corresponding literature for detailed explanations of these metrics.

As for the assertion classification task (Task B), we use the well-known Python package `sklearn`,[13] to calculate P, R and F1 scores. TP, FP and FN are computed per category $c \in$ {`absent` `posssible`} as follows:

- TP: number of medical entities of type c correctly classified as c.
- FP: number of medical entities of a type other than c incorrectly classified as c.
- FN: number of medical entities of type c incorrectly classified as a type other than c.

As average metrics of the different categories, we report micro-average scores (μ). The micro-average scores are obtained by applying the same Eqs. (1) to the sums of the TP, FP and FN of the different categories.

4. Results

4.1. Results of Task A: cue and scope detection

The main results of Task A, cue and scope detection, are shown in Table 7. We report per-category and micro-average F1-score results (see *SEM 2012 scores in Appendix C) of models trained in the full train set and one of the train subsets, with $\sim 1\%$ of examples.

Overall, we observe that the detection of cues (`NC` and `UC`) is easier than that of scopes (`NS` and `US`), and that speculation (`UC` and `US`) is more difficult to detected than negation (`NC` and `NS`). This is to be expected given the nature and distribution of each category, and was also noted by Lima-López et al. [8].

Regarding the differences among the systems trained on the full dataset, little difference among the Transformers is noted, although MarIA stands out with an average F1-score of 0.910, followed by BETO and XLM-RoBERTa (hereafter, XLM-R) – both 0.905 –. MarIA and XLM-R in particular achieve the greatest gains with respect to the uncertainty scope (`US`) scores of the baseline set by NCRF++, which presented the biggest opportunity for improvement in previous work. Unsurprisingly, SciBERT falls behind the other Transformers, but its performance is similar to Flair's. Still, both improve the baseline across all categories and manage to overpass prior state of the art [46] (see Table C.16 in Appendix C).

Looking at the performance of the models with the smaller train set, we see very significant gains of the Transformer models and Flair with respect to the baseline, particularly for uncertainty cues and scopes (`UC` and `US` respectively). It is remarkable that with only 169 examples of training, all the Transformer models yield F1-scores above 0.5 in the detection of uncertainty cues. It is noteworthy as well that the models that fare best with this smaller training set, BETO and IXAmBERT, are not the ones that achieve the best results when presented with the full training set. The behaviour of the models with increasing amounts of training data will be analysed in greater depth in a later Section 4.3.

4.2. Results of Task B: assertion classification

Table 8 shows the main results of Task B, assertion classification. Again, we report per-category and micro-average F1-results of models trained in the full train set and one of the train subsets (with $\sim 1\%$ of examples).

Similarly to Task A, MarIA obtains the best overall results (0.937 F1-score), followed by a multilingual model – mBERT in this case (0.935) – and BETO (0.934). Nevertheless, the differences between the Transformer models are narrower still than in the previous task, and even SciBERT manages to outperform some of the multilingual and

[13] https://scikit-learn.org

Table 7
F1-scores of Task A – cue and scope detection in the Full test set. The best and second-best scores are highlighted in bold and dotted underlines, respectively. N is the number of training examples.

	$1/3^4$ train set (N = 169)					Full train set (N = 13,802)				
	μ	NC	NS	UC	US	μ	NC	NS	UC	US
NCRF++	0.604	0.770	0.626	0.093	0.088	0.881	0.952	0.866	0.849	0.698
Flair + fastText	0.690	0.851	0.685	0.434	0.218	0.892	0.960	0.877	0.849	0.740
BETO	**0.735**	0.861	0.728	**0.616**	0.320	0.905	0.963	**0.900**	0.870	0.759
SpanBERTa	0.691	0.865	0.650	0.537	0.207	0.898	0.960	0.894	0.850	0.743
MarIA	0.708	0.855	0.699	0.529	0.283	**0.910**	**0.968**	0.897	**0.875**	**0.781**
IXAmBERT	0.730	0.854	**0.736**	0.609	0.322	0.901	0.965	0.888	0.865	0.755
mBERT	0.714	**0.866**	0.701	0.567	0.254	0.898	0.960	0.887	0.851	0.760
XLM-R	0.730	0.864	0.726	0.577	**0.324**	0.905	0.962	0.896	0.863	0.780
SciBERT	0.678	0.859	0.642	0.502	0.113	0.890	0.959	0.868	0.861	0.750

Table 8
F1-scores of Task B – assertion classification. The best and second-best scores are highlighted in bold and dotted underlines, respectively. N is the number of training examples.

	Full test set						Man test set		
	$1/3^4$ train set (N = 148)			Full train set (N = 12,108)			Full train set (N = 12,108)		
	μ	abs	pos	μ	abs	pos	μ	abs	pos
NegEx	0.647	0.698	**0.469**	0.683	0.700	0.638	0.890	0.922	0.783
Flair + fastText	0.003	0.004	0.000	0.889	0.892	0.882	0.939	0.951	0.903
BETO	0.612	0.729	0.409	0.934	**0.943**	0.914	0.972	0.979	0.952
SpanBERTa	**0.660**	0.759	0.330	0.927	0.937	0.905	0.967	0.971	0.955
MarIA	0.588	0.716	0.258	**0.937**	0.940	0.929	0.971	0.979	0.950
IXAmBERT	0.586	0.697	0.248	0.925	0.934	0.902	0.957	0.967	0.929
mBERT	0.635	0.731	0.438	0.935	0.939	0.925	0.973	0.978	**0.960**
XLM-R	0.647	**0.812**	0.292	0.934	0.934	**0.934**	**0.978**	**0.984**	0.959
SciBERT	0.458	0.586	0.149	0.927	0.931	0.916	0.967	0.975	0.943

Spanish models. The system based on Flair falls in average 3 F1-score points behind the worst Transformer.

All these systems outperform by far the baseline set by the rule-based system NegEx when allowed to exploit the whole training set, but mostly lag behind in the ∼ 1% training set scenario. Only SpanBERTa is capable of topping NegEx in this case. This issue will also be discussed in the next section.

In general, the task of assertion classification seems to be easier than cue and scope detection. The drop in performance from the negative class (absent) to the uncertainty class (possible) is also smaller.

It must be noted that none of the models with bigger vocabulary overlap with NUBes (i.e., SpanBERTa, IXAmBERT and BETO) nor the biggest models XLM-R, mBERT and IXAmBERT are absolute winners in either of the tasks. Although these models follow closely MarIA, none of the mentioned criteria seem to be decisive predictors of the performance of the models in these tasks.

4.3. Learning curves and adversarial examples

Regarding the learning curves for Task A (Fig. 7), NCRF++ shows the biggest gap between the scores for negation in the full test set and the rest of the scores along the whole curve, which evinces the greater capability of generalisation of the Transformer models and Flair.

It is striking that the most Spanish models set off with great advantage over the rest of the models where negation detection is concerned, although when looking at the scores for the most difficult examples, it becomes evident that all they are doing in practice is detecting the words "no" and "sin" ("without"). Given enough data, the other Transformers are capable of reaching the same performance quite quickly.

Most models (NCRF++, Flair, SciBERT and SpanBERTa most markedly) show an upwards trend still towards the end of the curve, which indicates they might be able to reach the results of the best models if given more data.

As for the learning curves of Task B (Fig. 8), we observe quite a different landscape. The gap between the full and harder test sets is much narrower than in the previous Task (except for NegEx), and the systems seem to reach a plateau with around a third of the training set. Furthermore, monolingual and multilingual models do not have such markedly different behaviours in this case. Most of all, Fig. 8 clearly demonstrates the problem of rule-based systems such as NegEx. Even if it is has a very good start at classifying the easiest negated instances, the system is just not capable of generalising to unseen cases even as the available data to enrich the tool lexicons increases.

4.4. Error analysis

We conclude the inspection of the results with an error analysis. We present the confusion matrices of the two tasks, and illustrate the most salient incorrect predictions.

For Task A (Fig. 9), the matrices have been computed at token level without taking into account the BIO tag of the tokens. The values are presented in relative terms ignoring true positive O predictions (being the majority class, it would render the matrices uninformative). That is, each matrix adds up to 1.

What the matrices show is that the most common errors are false negative errors of scopes, both for negation and speculation. The baseline NCRF++ is the system that commits this error more frequently, which accounts for ∼11% of its predictions (again, not considering the true Out tokens), while with BETO and XLM-R we manage to cut these errors by more than half. Still, the systems struggle to annotate scopes properly in the same contexts. We identified the following: sentences with coordination (Example 4.1), sentences with scopes preceding the cues (which typically involve relative clauses; Example 4.2), and sentences with negation or uncertainty reinforcement through multiple markers (4.3).[14]

(4.1). ***Ausencia de*** *factores de riesgo vascular, cardiopatía etc,* ..
"Absence of vascular risk factors, heart disease, etc".

[14] In the next examples, cues are highlighted in boldface and scopes in italics.

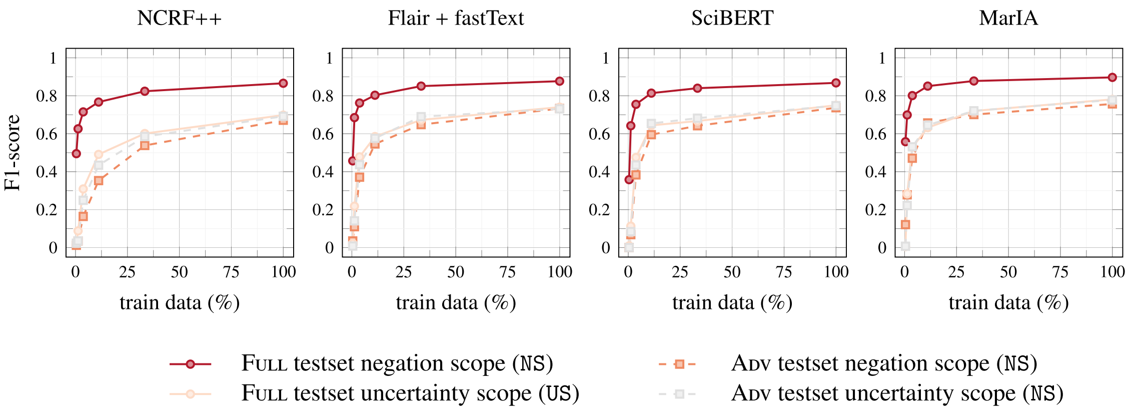


Fig. 7. Selected learning curves of Task A – cue and scope detection.

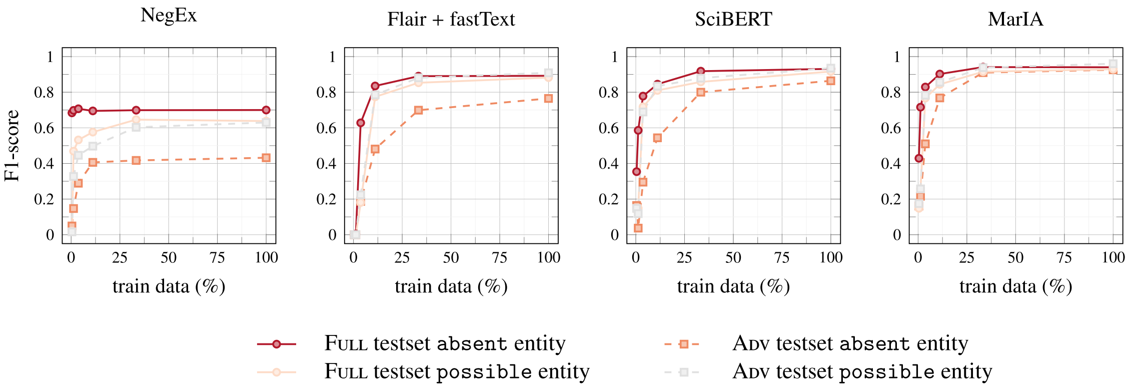


Fig. 8. Selected learning curves of Task B – assertion classification.

				predicted	
true	NC	NS	UC	US	O
NC	0.15	0.00	0.00	0.00	0.01
NS	0.00	0.47	0.00	0.01	0.06
UC	0.00	0.00	0.05	0.00	0.01
US	0.00	0.01	0.00	0.14	0.05
O	0.00	0.03	0.00	0.01	

NCRF++

				predicted	
true	NC	NS	UC	US	O
NC	0.15	0.00	0.00	0.00	0.00
NS	0.00	0.48	0.00	0.01	0.04
UC	0.00	0.00	0.05	0.00	0.00
US	0.00	0.01	0.00	0.16	0.03
O	0.01	0.03	0.00	0.02	

Flair + fastText

				predicted	
true	NC	NS	UC	US	O
NC	0.15	0.00	0.00	0.00	0.00
NS	0.00	0.49	0.00	0.01	0.03
UC	0.00	0.00	0.05	0.00	0.00
US	0.00	0.01	0.00	0.17	0.02
O	0.01	0.02	0.00	0.03	

BETO

Fig. 9. Selected confusion matrices of Task A – cue and scope detection.

(4.2). *[...] componente psiquiátrico añadido que* ***justificara*** *la crisis. '[...] an added psychiatric component that could justify the crises".*

(4.3). ***Intepreto*** *el cuadro clínico como* ***probable*** *pericarditis "I interpret the clinical picture as probable pericarditis"*

Although to a lesser extent, the systems make false positive errors as well when it comes to the detection of scopes. The most common of these errors stems from the inability of the systems to recognise as separate syntactic constituents a phrase or clause affected by negation/uncertainty and a following adjunct, as are "sobreinfectado" ("overinfected") and "en el lado derecho" ("on the right side") in Example 4.4:

(4.4). *Se observa hidrocele [...]* ***probablemente*** *sobreinfectado en lado dcho. "Probably overinfected hydrocele [...] observed on the right side".*

Even human annotators find these cases challenging, because the sentences may be syntactically ambiguous and must be interpreted mindfully to capture the physician's intended meaning in the annotations.

Finally, there seems to be a little confusion with some negation and speculation scopes among most systems: in ~1% of the tokens, some systems emit the tag US (uncertainty scope) when it should be NS (negation scope). Upon manual analysis of these cases, we consider that the systems are actually not committing errors but correcting what appear to be incorrect – or at best debatable – manual annotations, as exemplified in Table 9.

Regarding cues, some false negative errors involve infrequent lexical expressions that the systems were not able to generalise. This is particularly the case for uncertainty cues (UC). Here are a two examples undetected by the majority of the systems:

(4.5). ***Hay que asumir*** *que está infectada "It must be presumed that she is infected"*

(4.6). *Dice haber ingerido lorazepam [...] con ideación,* ***al parecer,*** *autolitica "They refer having ingested lorazepam [...] with apparent suicidal ideation"*

Further, a minor source of false negative cue annotations are errors caused by factors unrelated to the systems themselves, and that have

Table 9
Gold annotations and predictions on the sentence extract “unable to specify whether there was a loss of consciousness or not”. The fact that the phrase contains what are typically negative cues (“unable to”, “loss of”) and that the uncertainty cue is discontinuous (“whether [...] or not”) makes this example specially difficult to predict correctly. While the manual annotations interpret the phrase as a negation cue and scope, most of the systems (except Flair) retract their predictions midway in favour of speculation.

Token	Gold	NCRF++	MarIA	BETO	Flair
incapaz	B-NC	B-NC	B-NC	B-NC	B-NC
de	I-NC	I-NC	I-NC	I-NC	I-NC
precisar	B-NS	I-UC	B-NS	I-UC	B-NS
si	I-NS	I-UC	I-UC	I-UC	O
hubo	I-NS	I-UC	I-UC	B-US	O
o	I-NS	I-UC	B-UC	B-UC	O
no	I-NS	I-UC	I-UC	I-UC	B-NC
perdida	I-NS	B-US	B-US	B-US	B-NS
de	I-NS	I-US	I-US	I-US	I-NS
conocimiento	I-NS	I-US	I-NS	I-US	I-NS

to do with the limitations of the NUBes corpus. First, a few expressions are inconsistently annotated throughout the corpus, such as the verb “evitar” (“avoid”); the systems have learned not to interpret it as a negation cue, but it is annotated in the reference corpus in a minority of occurrences. Second, tokenisation errors in sentences with ungrammatical usage of punctuation marks induce errors in the post-processing of the predicted labels by the Transformers, as only the prediction of the first subword is taken as final label for a word. Take the following example:

(4.7). *Comenzar tolerancia oral.**Asintomática**. “Start oral tolerance. Asymptomatic”.*

While the systems may be able to detect properly that “Asintomática” (*asymptomatic*) is a negation cue, it will not be annotated as such because the word in the NUBes corpus is “oral. Asintomática” (sic) and only the label produced for the first subword (e.g., “oral”) is taken to account to produce the final labels.

In this case, the Flair sequence labeller produces the least false negative cues, missing out just 2% of the negation cues (NC) and 6% of the uncertainty cues (UC). NCRF++ is again the worst system, doubling the false prediction rates of Flair.

As for false positive predictions of cues, they actually stem for the most part from human errors, that is, these predictions capture cues overlooked by the human annotators. Interestingly, the error rates are inverted for this error set, with NCRF++ committing the least false positives and XLM-R leading the rank, followed closely by SpanBERTa. Pending an example-by-example manual revision, it seems sound to assume that XLM-R and SpanBERTa are not committing actual errors but simply detecting more human errors of the type just explained than the rest of the systems.

Regarding the confusion matrices of Task B (Fig. 10), false positive errors are much more frequent and, in fact, constitute the bulk of errors made by the systems overall. A manual analysis of these errors revealed that they primarily involve entities near cues but that are not in focus, as in the following examples (starred categories indicate that the predictions are incorrect):

(4.8). *No mejoró con la toma de <e> **Paracetamol** </e> . . . . . *abs “[The patient] did not improve with Paracetamol”.*

(4.9). *Cuadro confusional de probable caracter reactivo al <e> **proceso infeccioso** </e> *pos “Confusional state of probable reactive character to the infectious process”.*

(4.10). *Se aconseja TAC para valorar la causa de la <e> **obstrucción de la vía biliar** </e> . . *pos “CT is advised to assess the cause of bile duct obstruction”*

In Example 4.8, the focus of the negation is in the improvement of the patient, who *did* take Paracetamol. In Example 4.9, the relation between the confusional state and the infectious process is uncertain, not whether an infectious process took place (the use of determinate article “the” in “the infectious process” makes it clear that it is a reference to a known past event). Finally, in Example 4.10, what is unknown is the origin of the obstruction, not the existence of the obstruction itself (the same rationale applies here). These examples are particularly tricky because they require deeper understanding of the sentences than that needed to simply find cues and scopes. Even then, it is likely that fewer of this type of errors might occur if the models were trained on gold standard corpora instead of the automatically generated corpus described in this work.

As for false negative errors, we found two main types of instances that confuse the models:

1. Sentences that express a change of state, such as disappearance of symptoms or modifications in a treatment:

(4.11). *Presenta <e> **fiebre** </e> elevada que cede con tratamiento antibiótico . *pre “[The patient] has high fever that goes down with antibiotic treatment”.*

(4.12). *Le pautaron <e> **Diclofenaco** </e> que no está tomando *pre “[The patient] was prescribed Diclofenaco which she does not take”*

In these cases, the symptom or treatment is asserted in the main clause of the sentence but negated in the relative clause. Although debatable, the guidelines of the NUBes corpus indicate that these examples should be explicitly annotated as negations, but the models seem to struggle with such instances.

2. Long sentences where the scope precedes a negation cue, which occurs towards the end of the sentence:

(4.13). *Se obtienen muestras de <e> **cultivo de sangre** </a> y [...] siendo negativos . . *pre “Blood culture and [...] were obtained with negative result”.*

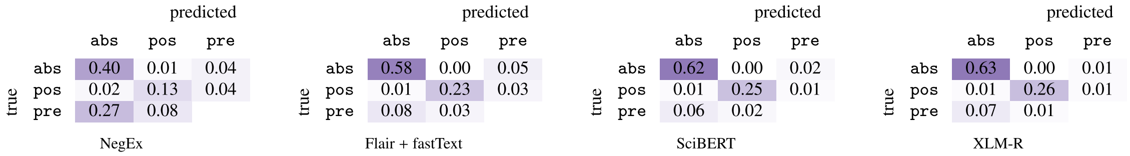
The long distance between the cue and the scope, as well as their less common order in the sentence, appears to make it more difficult for the systems to establish a relation between the two.

In the case of assertion classification, there does not seem to be much confusion between instances of negated and possible entities as there was in the cue and scope detection task.

Finally, as part of the error analysis, we studied whether the errors that the systems are making in the two separate tasks coincide somehow in the same examples, given that the corpora for the two tasks originate from the same collection of sentences. Out of the 2,762 sentences for testing in Task A, 272 have errors (made by any of the evaluated models). In Task B, the ratio is 196 out of 2,467. A significant amount, 92 sentences, are common to both tasks and involve most of the situations discussed in this section, with a prominent presence of sentences with relative clauses where scopes are discontinuous and may surround their cue.

5. Discussion

The presented results and their analysis lead to the discussion of which framework, assertion classification or cue and scope detection, is better suited and when. While the tasks’ results are not strictly comparable and the decision of using one framework over the other depends of course on the application objective of the system and the resources at hand to implement it, a few observations can be made on this subject:



predicted

true	abs	pos	pre
abs	0.40	0.01	0.04
pos	0.02	0.13	0.04
pre	0.27	0.08	

NegEx

predicted

true	abs	pos	pre
abs	0.58	0.00	0.05
pos	0.01	0.23	0.03
pre	0.08	0.03	

Flair + fastText

predicted

true	abs	pos	pre
abs	0.62	0.00	0.02
pos	0.01	0.25	0.01
pre	0.06	0.02	

SciBERT

predicted

true	abs	pos	pre
abs	0.63	0.00	0.01
pos	0.01	0.26	0.01
pre	0.07	0.01	

XLM-R

Fig. 10. Selected confusion matrices of Task B – assertion classification.

- First, the task of assertion classification may seem to be easier to learn, as systems reach a plateau with less training data, said plateau actually surpassing the best metrics obtained for the sequence labelling task. However, it must be acknowledged that the synthetic nature of the assertion classification corpus is likely playing a role in this regard. Further, as a trade-off, the entities to be classified must be annotated beforehand, which, if done automatically, would inject errors in the pipeline and potentially render this approach equally or more challenging than the detection of cues and scopes through sequence labelling.
- Similarly, the annotation effort could be said to be smaller for assertion classification: less data is required, and the annotation of assertion classification is easier for humans than that of cues and scopes. However, again, it requires that the corpus be annotated for the entities of interest. If such corpus does not exist, the manual annotation effort is increased substantially, unless automated methods are devised such as those described in this work (Section 3.1.2).
- While it may seem that a cue and scope sequence labeller can resolve the assertion classification task as a side-effect (by simply marking as negated or uncertain any entity under a detected scope), we have shown that this approach leads to false positives. At the same time, it is easier to convert a corpus annotated with cues and scopes to a corpus for assertion classification than the other way around.

Yet a third logical approach to the processing of negation and speculation could consist in an end-to-end assertion classification, that is, a sequence labeller for medical entities that would jointly detect the entities of interest and assign them an assertion category. This scenario exceeds the scope of this work and, to the best of our knowledge, has not been tested in the literature (closely related work exists [24,29,30] but it only focuses on detecting negated entities). Such a system would be more efficient than a pipeline composed of a medical entity recogniser and an assertion classifier, because it would solve the task in a single pass over the input text. The assertion classifier, in contrast, must be invoked as many times as entities to be categorised. On the other hand, in taking on the challenges of the two tasks, the assertion sequence labeller could require more and better data to achieve comparable results. Furthermore, the knowledge about negation and speculation captured by such a model would be harder to transfer to other domains, as it would be inextricably bound with the entity types the model was trained to detect.

Requiring a more elaborate architecture, multi-task learning offers another avenue of research. In this setup, the tasks of cue and scope detection and assertion classification would be learned jointly by the same model in separate classification heads, possibly benefiting one another. Interestingly, [45] find that learning to classify events into the affirmed or negated categories as an auxiliary task to negation scope resolution does not help and can even be detrimental. However, their setup exploits a different corpora per task and those corpora involve different languages. Furthermore, they do not look into how the task of negation scope resolution affects assertion classification.

Following the paradigm shift in the NLP community [72,73], future work may address the processing of negation and speculation with yet other emergent approaches, such as sequence-to-sequence and/or prompt-based learning, leveraging perhaps bigger language models (e.g., GPT3 [74], BART [75], T5 [76]). In this regard, while several works [77,78] demonstrate that language models are not good at capturing how negation changes the meaning the sentences they appear in, others [79,80] found evidence for some form of encoding of negation at the syntactic level (to the best of our knowledge, similar studies have not been conducted in regards to speculation). As the processing of negation and speculation, as addressed in this paper, is rather influenced by syntax than by semantics—i.e., the objective of the proposed systems is, in a nutshell, to decide *if*, not *how*, certain parts of a given sentence are affected by the presence of a negation or speculation cue—, these new paradigms may be found to be viable and even competitive for this task, as have been for others.

6. Conclusions

While the processing of negation has enjoyed much attention for years, the processing of speculation has not been studied to the same degree, most of all in languages other than English. In this work, we have evaluated multiple state-of-the-art models for sequence labelling and text classification in the tasks of negation and speculation cue and scope detection as well as assertion classification. The experiments have been conducted in a public corpus, NUBes [8], of health records written in Spanish. The evaluated systems include multiple BERT- and RoBERTa-like Transformer-based models, Flair, and two baseline systems.

The task of cue and scope detection was learned jointly by the systems. The Transformed-based model with the MarIA pre-trained model [68] achieved the best overall results (0.91 micro-average F1-score), advancing the state of the art previously set by Lima-López et al. [8] and Solarte Pabón et al. [46]. The system is closely followed by most of the other Transformer-based models, while SciBERT and the Flair sequence labeller fall slightly behind (still improving the baseline and previous state of the art). The improvement is brought predominantly by a better detection of speculation scopes as well as of the least frequent negation instances.

Regarding the assertion classification task, we first introduced an approach to convert the NUBes corpus, originally annotated for cues and scopes, to a corpus suitable for this task. A manual revision of the testing portion of the resulting corpus, as well as a manual error analyses of the results, suggest that this technique yields acceptable results and can be useful in scenarios were there is no such corpus available, as was the case in this work. In this task too, MarIA obtained the best results (0.937 micro-average F1-score), followed even more closely by the other Transformers, including SciBERT.

We observed that, in both tasks, neither the models with most vocabulary overlap with NUBes nor the biggest models obtained the best results, although they did follow closely MarIA. Further, the learning curves showed that, while monolingual Spanish models start off with certain advantage, being able to correctly emit predictions for the most frequent and repetitive instances, all the Transformer models manage to obtain similar results when allowed to exploit the entire training sets. The learning curves also suggested that less labelled data may be necessary for the assertion classification task than for the cue-scope detection task, although the results for the latter may be artificially inflated due to the corpus being synthetic.

A manual error analysis revealed that the most common errors in the cue and scope detection task are false negative errors involving

Table A.10
Dataset subsets for the training curves of Task A.

	$1/3^1$	$1/3^2$	$1/3^3$	$1/3^4$	$1/3^5$
Sentences	4,600	1,533	510	169	56
with negation	1,761	576	210	78	24
with uncertainty	386	127	44	16	6
with both	127	53	16	4	1
Negation cue	2,337	775	273	97	31
Negation scope	2,135	708	251	91	31
Uncertainty cue	586	212	67	24	11
Uncertainty scope	590	211	66	24	10
Total	5,648	1,906	657	236	83

Table A.11
Dataset subsets for the training curves of Task B.

	$1/3^1$	$1/3^2$	$1/3^3$	$1/3^4$	$1/3^5$
`negated`	782	277	92	34	11
`uncertain`	332	118	39	14	5
`present`	2,921	949	316	100	33
out of scope (OOS)	1,317	436	140	43	9
from assertion	1,604	513	176	57	24
Total	4,035	1,344	447	148	49

scopes, that is, the predicted scopes tend to fall short compared to the gold annotations. This is particularly true in relative clauses, where part of the scope of a cue might precede the cue. In the case of the assertion classification task, the most common errors involve false positive errors, where medical entities under the scope of cues but *not* in focus are incorrectly tagged as absent or possible instead of present. The manual error analysis also uncovered several incorrectly annotated instances, which will help us improve the quality of NUBes.

Finally, this study is limited to the most common paradigms when it comes to the processing of negation and speculation, and it focuses on a single corpus for analysis. Future work may explore new ways of addressing the problem, such as combining the two presented paradigms in a single multi-task architecture or adopting the rapidly evolving paradigm of prompt-based learning. Furthermore, future work should also address an extrinsic evaluation scenario to study whether the proposed method indeed benefits the downstream cases that rely on negation and uncertainty detection. Lastly, it is essential that future research validates these conclusions across different corpora to ensure broader applicability.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is partly supported by the projects DeepText (KK-2020-00088, SPRI, Basque Government, Spain) and DeepReading (RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE).

Appendix A. Data subset sizes for training curves

Tables A.10 and A.11 show, respectively, the size of each training subset used to compute the training curves of Task A (Fig. 7) and Task B (Fig. 8), respectively.

Appendix B. Hyperparameters

Tables B.12–B.14 report the hyperparameters of the neural sequence taggers and text classifiers. Values between squares brackets are options or ranges for the hyperparameter optimisation. Any hyperparameter not reported here takes the default value given by the corresponding training API.

Table B.12
Transformer sequence taggers and text classifiers.

Hyperparameter	Value
Pre-trained model	*see* Table 5
Batch size	8
Maximum input length	220
Optimiser	AdamW [81]
Learning rate	[1e-5 - 1e-4]
Warmup steps	[0 - 500]
Weight decay	[0.0 - 0.3]
Maximum epochs	30

Table B.13
Flair sequence tagger and text classifier.

Hyperparameter	Value
Pre-trained word emb.	BWES [61]
Pre-trained Flair emb.	`es-forward`, `es-backward`
bi-LSTM/GRU layers	1
Hidden dimensions	[128, 256]
Dropout rate	[0.0 - 0.5]
Batch size	[8, 16, 32]
Optimiser	SGD
Learning rate	[0.05 - 0.15]
Minimum learning rate	1e−4
Weight decay	[0.0 - 0.05]
Maximum epochs	60

Table B.14
NCRF++ sequence tagger (from [8]).

Hyperparameter	Value
Character emb. dimensions	30
Character CNN layers	1
Character hidden dimensions	50
Word emb. dimensions	300
Word bi-LSTM layers	1
Word hidden dimensions	200
Dropout rate	0.5
Batch size	16
Optimiser	SGD
Learning rate	0.005
L_2 regularisation	1e−8
Weight decay	0.001
Momentum	0
Maximum epochs	40

Appendix C. Additional metrics for cue and scope detection results

Table C.15 reports the performance of the sequence labelling models in terms of the metrics described by Morante and Blanco [39] for the *SEM 2012 shared task on resolving the scope and focus of negation, later also employed in the NEGES workshops [17,18], among others. The evaluation script is publicly available from the official website of the shared task.[15] Notice that the script is prepared to count one type of cues and one type of scopes (namely, negation cues and scopes). In order to report separate scores for negation and speculation, we post-processed the outputs of the systems to contain just negation or uncertainty predictions, then applied the evaluation script.

Table C.16 reports the performance of the sequence labellings models in terms of the metric described by Solarte Pabón et al. [46], to which we refer as "BIO-weighted token-level" scores throughout this paper.

[15] https://www.clips.uantwerpen.be/sem2012-st-neg

Table C.15

*SEM F1 scores of Task A – cue and scope detection in the Full test set. The best and second-best scores are highlighted in bold and dotted underlines, respectively. We refer the reader to Morante and Blanco [39] for an explanation of each metric. We include the results of Hartmann and Søgaard [45], who tackle the resolution of negation scopes. SU is a supervised Multilingual BERT model. ZS ST_{cat} refers to zero-shot performance of a Multilingual BERT model trained on the BioScope corpus [3] and the SFU Review Corpus [82].

	Negation						Uncertainty					
	Cues	Scopes			Glob	%CNS	Cues	Scopes			Glob	%CNS
		CM	NCM	Token				CM	NCM	Token		
NCRF++	94.68	88.38	88.85	90.51	88.67	81.54	84.68	75.39	75.60	75.52	75.00	64.41
Flair + fastText	95.38	89.49	90.01	91.58	89.38	83.22	85.71	77.89	78.69	78.67	77.83	69.71
BETO	95.78	90.86	91.76	**93.27**	90.88	85.42	86.44	80.32	81.07	81.62	80.32	**74.41**
SpanBERTa	95.50	90.63	91.37	92.81	90.57	84.81	85.19	78.18	78.97	79.86	77.92	70.59
MarIA	**96.31**	**91.42**	**92.03**	93.17	**91.48**	**85.78**	**86.72**	80.32	80.91	82.36	80.13	72.94
IXAmBERT	96.06	90.32	90.94	92.81	90.47	84.72	85.93	78.47	79.87	81.53	78.42	70.00
mBERT	95.49	90.62	91.20	92.51	90.66	84.89	86.19	78.83	79.80	79.31	78.64	71.47
XLM-R	95.77	90.98	91.66	93.24	90.97	85.42	86.58	**80.71**	**81.85**	**83.02**	**80.77**	74.12
SciBERT	95.40	89.05	89.74	91.83	89.21	82.51	86.19	77.83	78.83	79.58	77.65	70.00
[45] SU	–	–	–	95.66	–	–	–	–	–	–	–	–
[45] ZS ST_{cat}	–	–	–	90.24	–	–	–	–	–	–	–	–

Table C.16

BIO-tag weighted token-level scores (from Solarte Pabón et al. [46]) of Task A – cue and scope detection in the Full test set. The best and second-best scores are highlighted in bold and dotted underlines, respectively. In principle, the only difference between the Multilingual BERT (mBERT) model reported here and that of Solarte Pabón et al. [46] is the optimisation of some hyperparameters (see Section 3.2.5), whose impact is most noticeable for uncertainty scopes, the most challenging category of all.

	Negation						Uncertainty					
	Cues (NC)			Scopes (NS)			Cues (UC)			Scopes (US)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
NCRF++	0.95	0.94	0.95	0.93	0.88	0.90	0.87	0.82	0.85	0.84	0.69	0.76
Flair + fastText	0.95	**0.97**	0.96	0.92	0.90	0.91	0.84	0.87	0.86	0.80	0.79	0.79
BETO	0.95	**0.97**	0.96	0.94	**0.92**	**0.93**	0.86	0.88	0.87	0.80	0.84	0.82
SpanBERTa	0.95	**0.97**	0.96	0.94	0.91	0.92	0.85	0.87	0.86	0.80	0.82	0.81
MarIA	**0.97**	**0.97**	**0.97**	**0.95**	0.91	**0.93**	**0.88**	0.88	**0.88**	0.84	0.82	0.83
IXAmBERT	0.96	**0.97**	0.96	0.94	0.91	**0.93**	0.86	0.87	0.87	**0.85**	0.78	0.81
mBERT	0.96	0.96	0.96	0.94	0.91	0.92	0.86	0.87	0.86	0.80	0.81	0.81
mBERT [46]	0.95	0.93	0.95	0.90	0.86	0.88	0.86	0.83	0.84	0.75	0.70	0.72
XLM-R	0.95	**0.97**	0.96	0.93	**0.92**	**0.93**	0.85	**0.89**	0.87	**0.82**	**0.85**	**0.84**
SciBERT	0.95	0.96	0.96	0.92	0.91	0.91	0.86	0.88	0.87	0.81	0.81	0.81

References

[1] Laparra E, Mascio A, Velupillai S, Miller TA. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. Yearb Med Inform 2021;30(01):239–44.

[2] Morante R, Sporleder C. Modality and negation: An introduction to the special issue. Comput Linguist 2012;38(2):223–60.

[3] Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 2008;9(Suppl 11):S9.

[4] Dalianis H, Skeppstedt M. Creating and evaluating a consensus for negated and speculative words in a Swedish clinical corpus. In: Proceedings of the workshop on negation and speculation in natural language processing. Uppsala, Sweden: University of Antwerp; 2010, p. 5–13.

[5] Cruz Díaz NP, Morante Vallejo R, Maña López MJ, Mata Vázquez J, Parra Calderón CL. Annotating negation in spanish clinical texts. In: Proceedings of the workshop computational semantics beyond events and roles. Valencia, Spain: Association for Computational Linguistics; 2017, p. 53–8.

[6] Cheng K, Baldwin T, Verspoor K. Automatic negation and speculation detection in veterinary cinical text. In: Proceedings of the Australasian language technology association workshop 2017. 2017, p. 70–8.

[7] Marimon M, Vivaldi J, Bel N. Annotation of negation in the IULA Spanish clinical record corpus. In: Proceedings of the workshop computational semantics beyond events and roles. Valencia, Spain: Association for Computational Linguistics; 2017, p. 43–52.

[8] Lima-López S, Perez N, Cuadros M, Rigau G. NUBes: A corpus of negation and uncertainty in Spanish clinical texts. In: Proceedings of the 12th language resources and evaluation conference. Marseille, France: European Language Resources Association; 2020, p. 5772–81.

[9] Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than English: Opportunities and challenges. J Biomed Semant 2018;9:1–13.

[10] Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical Natural Language Processing: a methodical review. J Am Med Inform Assoc 2019;27(3):457–70.

[11] Fernández Vítores D. El español: Una lengua viva. Informe 2021. Technical report, Instituto Cervantes; 2021.

[12] Moreno Sandoval A, Salazar Garrote M. La anotación de la negación en un corpus escrito etiquetado sintácticamente. Revista Iberoamericana de Lingüística 2013;8:45–60.

[13] Campillos Llanos L, Martínez P, Segura-Bedmar I. A preliminary analysis of negation in a Spanish clinical records dataset. In: Actas del Taller de NEGación en ESpañol. 2017, p. 33–9.

[14] Cruz Díaz NP, Maña López MJ. Negation and speculation detection. Natural language processing, (no. 13). John Benjamins Publishing Company; 2019.

[15] Jiménez-Zafra SM, Morante R, Martín-Valdivia MT, Ureña-López LA. Corpora annotated with negation: An overview. Comput Linguist 2020;46(1):1–52.

[16] Morante R, Blanco E. Recent advances in processing negation. Nat Lang Eng 2021;27(2):121–30.

[17] Jiménez-Zafra SM, Cruz Díaz NP, Morante R, Martín-Valdivia MT. NEGES 2018: Workshop on negation in Spanish. Procesamiento del Lenguaje Natural 2018;62:21–8.

[18] Jiménez-Zafra SM, Cruz Díaz NP, Morante R, Martín-Valdivia M-T. NEGES 2019 task: negation in spanish. In: Proceedings of the Iberian languages evaluation forum (IberLEF 2019) co-located with 35th conference of the Spanish society for natural language processing (SEPLN 2019). 2019, p. 329–41, CEUR Workshop Proceedings.

[19] Jiménez-Zafra SM, Taulé M, Martín-Valdivia MT, Ureña-López LA, Martí MA. SFU review$_{SP}$-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. Lang Resour Eval 2018;52:533–69.

[20] Taulé M, Nofre M, González M, Martí MA. Focus of negation: Its identification in Spanish. Nat Lang Eng 2021;27(2):131–52.

[21] Bel-Enguix G, Gómez-Adorno H, Pimentel A, Ojeda-Trueba S-L, Aguilar-Vizuet B. Negation detection on Mexican Spanish tweets: The T-MexNeg corpus. Appl Sci 2021;11(9):1–22.

[22] Costumero R, López F, Gonzalo-Martín C, Millan M, Menasalvas E. An approach to detect negation on medical documents in Spanish. In: Brain informatics and health. Lecture notes in computer science, 8609, Warszawa, Poland: Springer; 2014, p. 366–75.

[23] Stricker V, Iacobacci I, Cotik V. Negated findings detection in radiology reports in Spanish: an adaptation of NegEx to Spanish. In: Workshop on replicability and reproducibility in natural language processing: Adaptative methods, resources and software At IJCAI 2015. 2015, p. 1–7.

[24] Santiso S, Casillas A, Pérez A, Oronoz M. Medical entity recognition and negation extraction: Assessment of NegEx on health records in Spanish. In: Bioinformatics and biomedical engineering. Lecture notes in computer science, vol. 10208, Granada, Spain: Springer; 2017, p. 177–88.

[25] Santamaría J. NegEx-MES. 2018, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

[26] Solarte-Pabón O, Menasalvas E, Rodriguez-González A. Spa-neg: An approach for negation detection in clinical text written in Spanish. In: Bioinformatics and biomedical engineering. Lecture notes in computer science, vol. 12108, Granada, Spain: Springer; 2020, p. 323–37.

[27] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301–10.

[28] Koza W, Filippo D, Cotik V, Stricker V, Muñoz M, Godoy N, et al. Automatic detection of negated findings in radiological reports for Spanish language: Methodology based on lexicon-grammatical information processing. J Digit Imaging 2019;32:19–29.

[29] Santiso S, Casillas A, Pérez A, Oronoz M. Word embeddings for negation detection in health records written in Spanish. Soft Comput 2019;23:10969–75.

[30] Santiso S, Pérez A, Casillas A, Oronoz M. Neural negated entity recognition in Spanish electronic health records. J Biomed Inform 2020;105:103419.

[31] Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th international conference on machine learning. Williamstown, MA, USA: Morgan Kaufmann; 2001, p. 282–9.

[32] Oronoz M, Gojenola K, Pérez A, Díaz De Ilarraza A, Casillas A. On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. J Biomed Inform 2015;56:318–32.

[33] Loharja H, Padró L, Turmo J. Negation cues detection using CRF on Spanish product review texts. In: Proceedings of NEGES 2018: Workshop on negation in Spanish co-located with the 34th SEPLN conference. 2018, p. 49–54, CEUR Workshop Proceedings.

[34] Fabregat H, Martinez-Romo J, Araujo L. Deep learning approach for negation cues detection in Spanish. In: Proceedings of NEGES 2018: Workshop on negation in Spanish co-located with the 34th SEPLN conference. 2018, p. 43–8, CEUR Workshop Proceedings.

[35] Fabregat H, Duque A, Martínez-Romo J, Araujo L. Extending a deep learning approach for negation cues detection in Spanish. In: Proceedings of the Iberian languages evaluation forum (IberLEF 2019) Co-Located with 35th conference of the Spanish society for natural language processing. 2019, p. 369–77, CEUR Workshop Proceedings.

[36] Beltrán J, González M. Detection of negation cues in Spanish: The CLiC-Neg system. In: Proceedings of the Iberian languages evaluation forum (IberLEF 2019) co-located with 35th conference of the Spanish society for natural language processing. 2019, p. 352–60, CEUR Workshop Proceedings.

[37] Domınguez-Mas L, Ronzano F, Furlong L. Supervised learning approaches to detect negation cues in Spanish reviews. In: Proceedings of the Iberian languages evaluation forum (IberLEF 2019) co-located with 35th conference of the Spanish society for natural language processing. 2019, p. 361–8, CEUR Workshop Proceedings.

[38] Giudice V. Aspie96 at NEGES (IberLEF 2019): negation cues detection in Spanish with character-level convolutional RNN and tokenization. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) Co-Located with 35th Conference of the Spanish Society for Natural Language Processing. 2019, p. 342–51, CEUR Workshop Proceedings.

[39] Morante R, Blanco E. *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. Montréal, Canada: Association for Computational Linguistics; 2012, p. 265–74.

[40] Sineva E, Grünewald S, Friedrich A, Kuhn J. Negation-instance based evaluation of end-to-end negation resolution. In: Proceedings of the 25th conference on computational natural language learning. Association for Computational Linguistics; 2021, p. 528–43.

[41] Jiménez-Zafra SM, Morante R, Blanco E, Martín Valdivia MT, Ureña López LA. Detecting negation cues and scopes in Spanish. In: Proceedings of the 12th language resources and evaluation conference. Marseille, France: European Language Resources Association; 2020, p. 6902–11.

[42] Shaitarova A, Furrer L, Rinaldi F. Cross-lingual transfer-learning approach to negation scope resolution. In: Proceedings of the 5th Swiss text analytics conference (SwissText 2020) & 16th conference on natural language processing (KONVENS 2020). 2020, p. 1–7, CEUR Workshop Proceedings.

[43] Shaitarova A, Rinaldi F. Negation typology and general representation models for cross-lingual zero-shot negation scope resolution in Russian, French, and Spanish. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics (NAACL 2021): Student research workshop. Association for Computational Linguistics; 2021, p. 15–23.

[44] Rivera Zavala R, Martínez P. The impact of pretrained language models on negation and speculation detection in cross-lingual medical text: Comparative study. JMIR Med Inform 2020;8(12):e18953.

[45] Hartmann M, Søgaard A. Multilingual negation scope resolution for clinical text. In: Proceedings of the 12th international workshop on health text mining and information analysis. Online: Association for Computational Linguistics; 2021, p. 7–18.

[46] Solarte Pabón O, Montenegro O, Torrente M, Rodríguez González A, Provencio M, Menasalvas E. Negation and uncertainty detection in clinical texts written in Spanish: a deep learning-based approach. PeerJ Comput Sci 2022;8:e913.

[47] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems. Long Beach, CA, USA: Curran Associates Inc.; 2017, p. 6000–10.

[48] Khandelwal A, Sawant S. NegBERT: A transfer learning approach for negation detection and scope resolution. In: Proceedings of the 12th language resources and evaluation conference. Marseille, France: European Language Resources Association; 2020, p. 5739–48.

[49] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics; 2020, p. 8440–51.

[50] Liu X, He P, Chen W, Gao J. Multi-task deep neural networks for natural language understanding. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Firenze, Italy: Association for Computational Linguistics; 2019, p. 4487–96.

[51] Dalloux C, Claveau V, Grabar N. Speculation and negation detection in French biomedical corpora. In: Proceedings of the international conference on recent advances in natural language processing. Varna, Bulgaria: INCOMA Ltd.; 2019, p. 223–32. http://dx.doi.org/10.26615/978-954-452-056-4_026.

[52] Ramshaw L, Marcus MP. Text chunking using transformation-based learning. In: Natural language processing using very large corpora. Text, speech and language technology, vol. 11, Springer; 1999, p. 157–76.

[53] Perez N, Cuadros M, Rigau G. Biomedical term normalization of EHRs with UMLS. In: Proceedings of the 11th international conference on language resources and evaluation. Miyazaki, Japan: European Language Resources Association; 2018, p. 2045–51.

[54] Perez N, Accuosto P, Bravo À, Cuadros M, Martínez-Garcia E, Saggion H, et al. Cross-lingual semantic annotation of biomedical literature: Experiments in Spanish and English. Bioinformatics 2020;36(6):1872–80.

[55] Bodenreider O. The unified medical language system (UMLS): Integrating biomedical terminology. Nucleic Acids Res 2004;32(Supplement 1):D267–70.

[56] Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. J Biomed Inform 2003;36(6):414–32.

[57] Yang J, Zhang Y. NCRF++: An open-source neural sequence labeling toolkit. In: Proceedings of ACL 2018, System Demonstrations. Melbourne, Australia: Association for Computational Linguistics; 2018, p. 74–9.

[58] Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R. FLAIR: An easy-to-use framework for state-of-the-art NLP. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (NAACL 2019): Demonstrations. Minneapolis, MN, USA: Association for Computational Linguistics; 2019, p. 54–9.

[59] Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling. In: Proceedings of the 27th international conference on computational linguistics. Santa Fe, NM, USA: Association for Computational Linguistics; 2018, p. 1638–49.

[60] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017;5:135–46.

[61] Soares F, Villegas M, Gonzalez-Agirre A, Krallinger M, Armengol-Estapé J. Medical word embeddings for Spanish: Development and evaluation. In: Proceedings of the 2nd clinical natural language processing workshop. Minneapolis, MN, USA: Association for Computational Linguistics; 2019, p. 124–33.

[62] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (NAACL 2019): Human language technologies, volume 1 (long and short papers). Minneapolis, MN, USA: Association for Computational Linguistics; 2019, p. 4171–86.

[63] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. 2019, arXiv:1907.11692.

[64] Cañete J, Chaperon G, Fuentes R, Pérez J. Spanish pre-trained BERT model and evaluation data. In: Proceedings of the practical ML for developing countries workshop (PML4DC 2020) At the 8th international conference on learning representations (ICLR 2020). 2020, p. 1–9.

[65] Otegi A, Agirre A, Campos JA, Soroa A, Agirre E. Conversational question answering in low resource scenarios: A dataset and case study for Basque. In: Proceedings of the 12th language resources and evaluation conference. Marseille, France: European Language Resources Association; 2020, p. 436–42.

[66] Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing. Hong Kong, China: Association for Computational Linguistics; 2019, p. 3615–20.

[67] Ortiz Suárez PJ, Romary L, Sagot B. A monolingual approach to contextualized word embeddings for mid-resource languages. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Online: Association for Computational Linguistics; 2020, p. 1703–14.

[68] Gutiérrez-Fandiño A, Armengol-Estapé J, Pàmies M, Llop-Palao J, Silveira-Ocampo J, Carrino CP, et al. MarIA: Spanish language models. Procesamiento del Lenguaje Natural 2022;68:39–60.

[69] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-art natural language processing. 2019, arXiv:1910.03771.

[70] Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: A research platform for distributed model selection and training. 2018, arXiv:1807.05118.

[71] Bergstra J, Yamins D, Cox DD. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: Proceedings of the 30th international conference on machine learning, vol. 28. Atlanta, GA, USA: JMLR.org; 2013, p. 115–23.

[72] Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: A systematic survey of prompting methods in Natural Language Processing. 2021, arXiv:2107.13586.

[73] Sun T, Liu X, Qiu X, Huang X. Paradigm shift in natural language processing. 2021, arXiv:2109.12575.

[74] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Advances in Neural Information Processing Systems 33 (NeurIPS 33). Vancouver, Canada: Curran Associates, Inc.; 2020, p. 1877–901.

[75] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics; 2020, p. 7871–80.

[76] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2020;21(140):1–67.

[77] Kassner N, Schütze H. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Online: Association for Computational Linguistics; 2020, p. 7811–8.

[78] Ettinger A. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Trans Assoc Comput Linguist 2020;8:34–48.

[79] Warstadt A, Cao Y, Grosu I, Peng W, Blix H, Nie Y, et al. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing. Hong Kong, China: Association for Computational Linguistics; 2019, p. 2877–87.

[80] Zhao Y, Bethard S. How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics; 2020, p. 4729–47.

[81] Loshchilov I, Hutter F. Decoupled weight decay regularization. In: Proceedings of the 7th international conference on learning representations. 2019, p. 1–18.

[82] Konstantinova N, de Sousa SC, Cruz NP, Maña MJ, Taboada M, Mitkov R. A review corpus annotated for negation, speculation and their scope. In: Proceedings of the eighth international conference on language resources and evaluation. Istanbul, Turkey: European Language Resources Association (ELRA); 2012, p. 3190–5.