Negation and Uncertainty Detection using Classical and Machine Learning Techniques

Piotr Bonar[a], **Iker Romero Cespedes**[b], **Miriam Morales Franco**[c], **Suzana Jeal**[d] **and Adnan Boukfal Lazaar**[e]

[a] *1759684*
[b] *1635239*
[c] *1706106*
[d] *1707160*
[e] *1607081*

Abstract—In this project, we explore the task of negation and uncertainty detection in clinical narratives, an important challenge in the field of Natural Language Processing (NLP) within healthcare. Clinical documents often contain statements that express negation (e.g., "no evidence of pneumonia") or uncertainty (e.g., "possible infection"), and accurately identifying these cues is crucial for clinical decision-making, automated coding, and data extraction.

Keywords—*negation, uncertainty, detection, clinical narratives, Natural Language Processing, rule-based system*

Contents

1. Introduction

In this project, we address the task of negation and uncertainty detection in clinical narratives, a subfield of Natural Language Processing (NLP) that is particularly significant in the healthcare domain. Clinical records often contain statements that negate a condition (e.g., "no evidence of pneumonia") or express uncertainty (e.g., "possible infection"). Properly identifying these linguistic phenomena is critical for clinical decision-making, automated coding, data extraction for research, and decision-support systems.

Negation and uncertainty detection is typically framed as a sequence labeling problem, where the goal is to identify specific cue expressions (e.g., "no", "niega", "podría") and delineate the scope of these cues — that is, the span of text that they semantx'ically affect. The annotated output labels used in this project are:

- NEG: Negation cue (e.g., "no", "niega")
- NSCO: Scope of the negation
- UNC: Uncertainty cue (e.g., "posible", "sugiere")
- USCO: Scope of the uncertainty

This is a challenging task due to the nature of clinical texts, which are typically unstructured, use domain-specific abbreviations, include multiple languages (Catalan and Spanish), and often lack grammatical completeness. Furthermore, statements in clinical documents may mix positive, negative, and speculative assertions within the same paragraph, increasing the complexity of scope identification.

The project is structured in three stages:

- A rule-based system (focus of this first deliverable)
- A machine learning system (next deliverable)
- An optional deep learning model (for additional credit and comparison)

In this first deliverable, we are asked to build a thorough understanding of the task and the dataset, review relevant literature, and implement a strong rule-based baseline that can accurately identify cues and their scopes in medical text. To do this, we will start with exploring and analysing the data.

1.1. Data Exploration and Analysis

We are provided with two main datasets:

- `negacio_train_v2024.json`: This is the labeled training dataset.
- `negacio_test_v2024.json`: This is the test dataset, which may contain either empty annotations (for evaluation) or system predictions.

Each file contains a list of clinical discharge reports represented in JSON format. Each entry corresponds to one patient report and contains:

- A field `data.text` with the full textual content of the clinical document.
- A field `annotations` with a list of labeled spans (when available).

Each annotation includes:

- start and end character offsets in the text field.
- A labels list with one of the four possible tags: NEG, NSCO, UNC, USCO.

1.2. Structure and Format of the Text

The clinical texts have the following characteristics:

- Free-form style: Notes are written in telegraphic, non-standard grammar. Sentences may lack verbs or punctuation.
- Redacted information: Patient names and identifiers are replaced with placeholders like `** *** ***` or `(********)`.
- Multilingual: Texts mix Catalan and Spanish within the same document. This introduces lexical and syntactic variation.
- Domain-specific language: Includes medical jargon (e.g., "amniorrexis", "PEG", "duda diagnóstica"), abbreviations, and technical terms.
- Numerical data and metadata: Some sections include lab results or administrative data (e.g., "glucosa 103 mg/dL", "edad: 42 años").

Example excerpt from a report:

"niega dolor torácico, disnea, tos o fiebre."

Cue: `niega` → labeled as NEG
Scope: `dolor torácico, disnea, tos o fiebre` → labeled as NSCO

1.3. Challenges in the Data

- **Multilingual content**: Requires handling both Catalan and Spanish tokens, grammar, and cues.
- **Noisy formatting**: Redacted placeholders and inconsistent punctuation can break tokenization.
- **Free text structure**: Fragmented, ungrammatical phrases are common.
- **Ambiguity**: Words like "sin", "podría", "duda" can be either cues or neutral, depending on context.
- **Variable scope length**: Scope may be a noun phrase or a long subordinate clause.
- **Redundancy**: Some phrases may repeat similar concepts, complicating scope resolution.

1.4. Quantitative Observations (from sample inspection)

- The average report length ranges from 300 to 2,000+ characters.
- Most documents include multiple cue expressions, usually 2–10 per document.
- Scopes vary from 3-word noun phrases to full sentences.
- Many examples contain both negation and uncertainty simultaneously.

1.5. Data Requirements for Preprocessing

To prepare the text for analysis and tagging, we must:

- Normalize Unicode and remove excess whitespace.
- Handle special tokens (e.g., redacted identifiers).
- Use multilingual tokenization with a library like spaCy or Stanza.
- Develop and maintain separate cue lists for negation and uncertainty.
- Maintain robust matching logic for cue + scope alignment.

This preprocessing pipeline will serve both our rule-based baseline and future ML models.

2. Rule-Based Approaches: Literature Review

Rule-based systems are among the earliest and most interpretable methods for detecting negation and speculation in biomedical texts. One of the most well-known approaches is NegEx, developed by Chapman et al. (2001), which identifies negated concepts in clinical documents using a simple yet effective pattern-matching algorithm.

NegEx uses:

- A predefined lexicon of negation triggers (e.g., "no", "denies", "without").
- A regular expression pattern to define the direction of scope: either forward (cue precedes the scope) or backward (cue follows the scope).
- A fixed-size window (typically 5 tokens) around the cue word to define the span.

2.1. Extensions and Variants

Several enhancements have been proposed over the years:

- **ConText (Harkema et al., 2009)**: Adds support for uncertainty, experiencer (e.g., family history vs. patient), and temporality.
- **pyConTextNLP**: Python implementation with customizable modifiers and targets.
- **NegBio**: Uses dependency parsing for improved scope detection in radiology reports.
- **NegTool (Enger et al., 2021)**: Open-source toolkit with support for uncertainty and multilingual documents.

2.2. Limitations of Rule-Based Systems

- **Low recall**: Misses new or uncommon cue expressions not in the lexicon.
- **Static**: Requires manual updates when language or style changes.
- **Ambiguity**: Cannot resolve syntactic ambiguity or subtle semantic cues.
- **No context awareness**: Lacks ability to use long-range dependencies or word meaning.

Despite these limitations, rule-based systems remain highly effective as baselines, especially in domains like medicine, where vocabulary is limited and high precision is important.

2.3. Relevance to Our Project

Our dataset shares similarities with those used in NegEx and ConText research: clinical notes, negation and uncertainty, short contexts. We will build a rule-based baseline inspired by NegEx:

- Construct separate lexicons for negation and uncertainty.
- Use regular expressions and heuristics to extract scopes (within a 5-token window or until a punctuation/conjunction).
- Evaluate manually and explore edge cases.

This will provide a solid foundation before implementing more complex learning-based approaches in subsequent phases of the project.

3. Rule-Based Approach: Implementation

In this section, we describe the implementation of our rule-based system designed to detect negation and uncertainty in clinical texts written in Spanish and Catalan. The system identifies cue expressions and their respective scopes by applying handcrafted linguistic rules, offering both transparency and adaptability.

3.1. Overview

The system operates by scanning clinical documents for known negation (NEG) and uncertainty (UNC) triggers. Upon detection, it assigns two types of annotations: the trigger itself (NEG or UNC) and its corresponding scope (NSCO or USCO). A fixed-size token window is used to determine the scope of influence surrounding each trigger.

3.2. Trigger Lexicons

Two curated lists of trigger phrases were used:

- **Negation Triggers (NEG):** *"no", "sin", "niega", "niegan", "negado", "negativa", "no hay evidencia de", "ausencia de", "descarta", "descartan", "no presenta", "negativo", "negativos", "negativas", "neg", "afebril"*
- **Uncertainty Triggers (UNC):** *"posible", "probable", "sugiere", "sospecha", "podría", "aparentemente", "dudoso", "no se puede descartar", "es posible que", "probablemente", "sugestiva de", "sugestivo de", "dudosa", "se orienta", "valorar", "podria", "se considera", "no se descarta", "puede ser", "sugiere que", "podría tratarse de", "no se puede excluir", "no se descartan", "compatible con", "no concluyente", "sugiere diagnóstico de", "pudiera corresponder a", "puede representar", "debería considerarse", "presuntivamente", "hallazgos no concluyentes", "aspecto compatible con", "sospecha de", "compatibles con", "sugestivos de", "parece", "aparente", "sugestivas de", "posiblemente", "probables", "sospechosa de", "dudosamente", "impresiona", "desconoce", "posibilidad de", "no se puede asegurar", "difícil valorar", "hallazgos ambiguos", "aspecto que podría corresponder", "probabilidad baja de", "no se puede confirmar ni descartar", "sin signos concluyentes", "sospechoso de"*

These lists support both single-word and multi-word expressions, and were iteratively expanded using examples from annotated corpora.

3.3. Text Preprocessing

Before processing, the text undergoes normalization:

- Lowercasing all tokens.
- Removing accents using Unicode normalization.
- Stripping leading/trailing punctuation.

Sentences are segmented using punctuation-based rules, and then tokenized on whitespace. This ensures consistency across varying styles of clinical note-taking.

3.4. Scope Detection Logic

Each sentence is scanned for trigger expressions. When a match is found, the system defines a scope window of `WINDOW_SIZE=5` tokens before and after the trigger. Two annotations are recorded:

1. The trigger itself (labeled as NEG or UNC).
2. The scope of the trigger (labeled as NSCO or USCO).

This allows the system to distinguish between a negated condition (e.g., "no presenta fiebre") and an uncertain one (e.g., "posible infección").

3.5. Character-Level Annotation

For each detected span, the system calculates character-level offsets to enable integration with annotation tools. This includes careful handling of spacing to ensure accurate mapping from tokens back to raw text.

3.6. Dataset-Wide Application

The complete rule-based pipeline is encapsulated in a function that processes all documents in a JSON dataset. For each text entry, predicted annotations are appended in a compatible format for evaluation and visualization.

3.7. Benefits and Limitations

The main advantage of this approach lies in its transparency and robustness in low-resource settings. However, it may miss nuanced expressions outside the predefined trigger lists or fail when sentence structure is irregular. These cases could be mitigated by extending the rule set or combining it with learning-based methods in the future.

4. Rule-Based System: Evaluation

To assess the effectiveness of our rule-based approach, we developed a custom evaluation function tailored to the structure of our annotated dataset. The objective was to compare system-generated annotations against gold standard annotations and compute standard evaluation metrics: **precision**, **recall**, and **F1-score**, separately for each annotation label (**NEG**, **NSCO**, **UNC**, **USCO**).

4.1. Evaluation Methodology

The evaluation was conducted in three stages:

1. **Span Extraction:** For each document, predicted and gold-standard annotations were parsed to extract labeled spans defined by start and end character offsets.
2. **Matching Strategy:** A predicted span was counted as a true positive if it overlapped with a gold span of the same label that had not yet been matched. Unmatched predictions were counted as false positives, and unmatched gold spans were counted as false negatives.
3. **Metric Calculation:** Standard metrics were calculated for each label:
   - **Precision** = TP / (TP + FP)
   - **Recall** = TP / (TP + FN)
   - **F1-Score** = 2 · (Precision · Recall) / (Precision + Recall)

This approach prioritizes partial span overlaps rather than requiring exact matches, which is more realistic for scope-based annotation tasks.

4.2. Evaluation Results and Analysis

The system was evaluated on the annotated training dataset, yielding the following results:

Label	Precision	Recall	F1-Score
NEG	0.94	0.92	0.93
NSCO	0.92	0.95	0.93
UNC	0.73	0.82	0.77
USCO	0.75	0.84	0.79

Table 1. Evaluation scores for each annotation label.

- **High performance in negation detection:** The system achieved strong precision and recall for **NEG** and **NSCO**, reflecting its accuracy and consistency in identifying negation triggers and scopes.
- **Moderate performance in uncertainty detection:** While performance for **UNC** and **USCO** was lower, particularly in recall, iterative refinement of the trigger lists helped improve coverage over time.
- **Strong overall precision:** Thanks to normalization techniques (e.g., removing accents and punctuation) and including a wide variety of negation and uncertainty expressions we were able to improve the coverage

These results support the viability of transparent rule-based systems in low-resource and high-accountability environments such as healthcare, especially for structured tasks like negation and uncertainty detection.