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 **1**
1.1 Data Exploration and Analysis . . . 1
1.2 Structure and Format of the Text . . . 2
1.3 Challenges in the Data . . . 2
1.4 Quantitative Observations (from sample inspection) . . . 2
1.5 Data Requirements for Preprocessing . . . 2
2 Rule-Based Approaches: Literature Review **2**
2.1 Extensions and Variants . . . 2
2.2 Limitations of Rule-Based Systems . . . 2
2.3 Relevance to Our Project . . . 2
3 Rule-Based Approach: Implementation **2**
3.1 Overview . . . 2
3.2 Trigger Lexicons . . . 3
3.3 Text Preprocessing . . . 3
3.4 Scope Detection Logic . . . 3
3.5 Character-Level Annotation . . . 3
3.6 Dataset-Wide Application . . . 3
3.7 Benefits and Limitations . . . 3
4 Rule-Based System: Evaluation **3**
4.1 Evaluation Methodology . . . 3
4.2 Evaluation Results and Analysis . . . 3
5 Machine Learning: Literature Review **3**
5.1 CRF-Based Approaches . . . 4
5.2 Lexicon-Filtered SVM: NegTool . . . 4
5.3 Meta-Learning for Scope Detection . . . 4
5.4 Summary for Implementation . . . 4
6 Machine Learning: Implementation **4**
6.1 Overview . . . 4
6.2 Data Input and Structure . . . 4
6.3 Preprocessing and Tokenization . . . 4
6.4 BIO Label Generation . . . 4
6.5 Feature Extraction . . . 4
6.6 Vectorization and Label Encoding . . . 5
6.7 Model Training and Evaluation . . . 5
6.8 Conclusion . . . 5
7 Machine Learning: Evaluation **5**
7.1 Evaluation Results . . . 5
7.2 Performance Analysis . . . 5
7.3 Averaged Metrics . . . 5
7.4 Conclusion . . . 5
8 Summary and Comparison **5**
8.1 Rule-Based Approach . . . 5
8.2 Machine Learning Approach . . . 5
8.3 Final Comparison and Outlook . . . 6

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)

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.

5. Machine Learning: Literature Review

This section summarizes four relevant works that guided the design of our machine learning system for negation cue detection in Spanish clinical texts.

5.1. CRF-Based Approaches

Loharja et al. (2018) and **CLiC-Neg** (2019) applied **Conditional Random Fields (CRF)** with BIO tagging for cue detection. Their models leveraged rich feature sets, including lexical information (token and lemma), part-of-speech (POS) tags, orthographic features, affixes, and wide context windows (up to six tokens before and one after). These CRF systems achieved high F1 scores (84–88%), significantly outperforming rule-based baselines.

Although both teams experimented with handcrafted rules and cue lexicons, these additions offered minimal improvement or even harmed performance. This highlights that, when enough annotated data and well-engineered features are available, data-driven CRF models are more effective and adaptable than rule-based systems.

Takeaway: CRF is a strong baseline for cue detection in Spanish. Feature engineering is more beneficial than hybrid rule-based additions.

5.2. Lexicon-Filtered SVM: NegTool

Enger et al. (2017) presented **NegTool**, a lightweight open-source system for detecting negations in English. They used a **Support Vector Machine (SVM)** classifier for cue detection, restricted to candidate words filtered through a predefined lexicon. Despite this simplification, their system achieved an F1 of 91.8%, ranking among the top performers in the SEM 2012 task. Scope detection was handled using a max-margin CRF with syntactic features.

NegTool's approach demonstrates that lexicon-filtered SVMs can be efficient and accurate, particularly in resource-constrained settings or when the vocabulary is stable.

Takeaway: Lexicon-filtered SVMs offer a practical alternative to CRFs, trading off some flexibility for speed and simplicity.

5.3. Meta-Learning for Scope Detection

Morante and Daelemans (2009) focused on scope detection, using a two-step system: cue detection with IGTREE (a decision tree-based learner), followed by scope prediction using three separate classifiers (CRF, SVM, TiMBL) whose outputs were combined in a CRF meta-learner. Their method yielded strong scope-level accuracy, particularly in biomedical texts.

Although relevant for future development, this system is beyond the scope of our current deliverable, which focuses only on cue detection.

5.4. Summary for Implementation

- We adopt a CRF model with BIO tagging, using features inspired by Loharja et al. and CLiC-Neg.
- As an alternative, we may experiment with a lexicon-filtered SVM similar to NegTool.
- Rule-based corrections will be used cautiously, only for well-identified systematic errors.
- Scope detection and meta-learning are deferred for future work.

This strategy builds on proven architectures while tailoring the system to the specific challenges of negation cue detection in Spanish.

6. Machine Learning: Implementation

In this section, we describe the implementation of our machine learning-based system to detect negation and uncertainty cues and scopes in clinical narratives. The objective is to build a lightweight and transparent model that predicts BIO-formatted labels (`B-NEG`, `I-USCO`, etc.) Unlike the rule-based system, which relies on handcrafted linguistic patterns and predefined cue lexicons, the machine learning model is trained independently from annotated data and does not incorporate any prior rule-based output. It offers an alternative, data-driven approach to the task of cue and scope detection.

6.1. Overview

The implemented machine learning system processes clinical documents by first splitting each text into tokens and aligning them with annotated spans to generate BIO-formatted labels. For every token, a set of simple lexical and contextual features is extracted, capturing characteristics such as casing, digit composition, and neighboring words. These features are then transformed into a numerical representation suitable for machine learning using vectorization techniques. The corresponding labels are encoded, and the data is split into training and test sets. A linear classifier is trained to predict the label of each token based on its features, allowing the system to identify cues and scopes of negation and uncertainty directly from the data without relying on predefined rules or lexicons.

6.2. Data Input and Structure

The input dataset consists of clinical discharge reports, where each entry includes:

- A raw text field located under `data.text`, containing the full clinical narrative.
- Manually annotated spans found under `predictions[0].result`, where each annotation provides the start and end character offsets of the span within the text and a label from a fixed set: `NEG`, `NSCO`, `UNC`, or `USCO`.

These annotations are used to reconstruct token-level labels by aligning character-level spans with token boundaries. Each token is assigned a BIO tag such as `B-NEG`, `I-USCO`, or `O` depending on its position relative to the annotation span.

6.3. Preprocessing and Tokenization

Each document is pre-processed using a tokenization strategy based on whitespace. Punctuation characters are removed from the text prior to token splitting. Although this method is relatively simplistic, it provides high efficiency and facilitates consistent offset tracking between tokens and annotated spans, which is crucial for accurate label assignment.

6.4. BIO Label Generation

To map span-level annotations to individual tokens, the system compares each token's character position with the provided annotation offsets. When a token overlaps with a span, it is assigned a label using the BIO scheme:

- The first token in a span receives a `B-` (Beginning) label.
- Subsequent tokens in the same span receive `I-` (Inside) labels.
- The tokens outside any span are labeled as `O` (Outside).

This transformation is essential to convert span-level annotations into a format suitable for token-level classification.

6.5. Feature Extraction

For each token, a compact yet informative feature vector is constructed. The features include:

- **Lexical features:** lowercased token form, uppercase status, titlecase status, and digit presence.
- **Contextual features:** surface forms of the previous and next tokens (if available).
- **Positional features:** binary flags indicating whether the token is at the beginning or end of the sentence (BOS/EOS).

These features are designed to capture basic lexical and syntactic patterns commonly associated with cue and scope expressions in clinical text.

6.6. Vectorization and Label Encoding

Once all features are extracted, they are transformed into numerical vectors using `DictVectorizer`, which supports sparse matrix representation for memory efficiency. Corresponding BIO labels are encoded into integer form using `LabelEncoder` to prepare the data for classification.

6.7. Model Training and Evaluation

The dataset is randomly split into training and testing subsets using an 80/20 ratio. The classifier used is `SGDClassifier` from the `scikit-learn` library, which applies stochastic gradient descent optimization. This model is particularly well-suited for high-dimensional sparse feature spaces and offers fast training times.

After training, the model is evaluated using the `classification_report` function, which provides precision, recall, and F1-score for each label. This evaluation allows for detailed analysis of model performance across different cue and scope types.

6.8. Conclusion

This implementation provides a transparent and modular machine learning pipeline for negation and uncertainty detection. While simple in architecture, it establishes a strong baseline for comparison with more sophisticated models such as Conditional Random Fields (CRFs) or neural architectures. The system also demonstrates how span-based clinical annotations can be effectively converted into token-level labels for use in supervised learning tasks.

7. Machine Learning: Evaluation

To evaluate the effectiveness of the implemented machine learning model, we evaluated its predictions on a held-out test set using standard classification metrics: precision, recall, F1-score, and support (i.e., the number of true instances for each class). The evaluation was conducted using the `classification_report` function from `scikit-learn`, which provides per-class statistics as well as macro and weighted averages.

7.1. Evaluation Results

Label	Precision	Recall	F1-Score	Support
B-NEG	0.96	0.95	0.95	371
B-NSCO	0.94	0.85	0.89	354
B-UNC	0.90	0.41	0.57	46
B-USCO	0.83	0.39	0.53	49
I-NEG	1.00	0.50	0.67	6
I-NSCO	0.83	0.39	0.53	702
I-UNC	0.88	0.68	0.77	22
I-USCO	1.00	0.01	0.03	140
O	0.96	1.00	0.98	15291

Table 2. Evaluation metrics for each BIO label on the test set.

The model achieved an overall accuracy of 95%. It performs strongly on common labels, especially for negation cues and scopes, while struggling with lower-frequency and longer-span labels.

7.2. Performance Analysis

- **Negation detection:** The labels `B-NEG` and `B-NSCO` yielded high precision and recall, confirming the model's abilities in identifying negation cues and their immediate scope.
- **Uncertainty detection:** Labels such as `B-UNC` and `B-USCO` achieved good precision but much lower recall, suggesting the model often misses uncertain expressions or fails to detect them fully.
- **Low recall for I-labels:** While `I-NEG`, `I-UNC`, and `I-NSCO` perform moderately well given their limited support, the `I-USCO` label stands out with a precision of 1.00 but a recall of only 0.01. This indicates severe under-detection of multi-token uncertainty scopes.
- **High precision on non-relevant tokens:** The `O` label—representing tokens not part of any annotation—was predicted with very high accuracy. This improves overall precision but also reflects the class imbalance, as it dominates the dataset with 15,291 instances.

7.3. Averaged Metrics

- **Macro average:** Precision = 0.92, Recall = 0.58, F1-score = 0.66. This unweighted average reflects the model's difficulty with low-frequency and complex labels.
- **Weighted average:** Precision = 0.95, Recall = 0.95, F1-score = 0.94. These scores are influenced by the model's strong performance on the majority `O` class and more frequent `B-*` labels.

7.4. Conclusion

These results suggest that a simple feature-based model can perform well on core negation detection tasks in clinical text. However, longer or more context-sensitive scopes, particularly for uncertainty, remain difficult to capture with local token-level features alone. Future improvements may include sequence-aware models (e.g., CRFs) or transformer-based language models to better handle complex and less frequent patterns.

8. Summary and Comparison

In this project, we addressed the task of detecting negation and uncertainty in Spanish and Catalan clinical narratives using two main approaches: a rule-based system and a machine learning model. Both systems were developed, evaluated, and compared based on their ability to correctly identify cues (`NEG`, `UNC`) and their scopes (`NSCO`, `USCO`).

8.1. Rule-Based Approach

The rule-based method, inspired by NegEx and ConText, relied on carefully curated lexicons of trigger phrases and a fixed-size window-based scope detection heuristic. It achieved strong performance in negation detection, particularly for the `NEG` and `NSCO` labels, with F1-scores around **0.93**. However, the system performed moderately in detecting uncertainty cues and scopes, where the F1-scores dropped to **0.77** and **0.79**.

Strengths:

- Transparent and interpretable
- High precision for structured negation
- Robust in low-resource environments

Limitations:

- Static and brittle with respect to language variability
- Limited recall due to fixed trigger lists
- Difficulty in capturing subtle or non-standard expressions

8.2. Machine Learning Approach

The machine learning model used a simple linear classifier (`SGD`) trained on lexical and contextual token-level features in a `BIO-tagging` format. The model showed high accuracy for the more frequent classes (`B-NEG`, `B-NSCO`, and `O`) with F1-scores up to **0.95**, but struggled with rare and longer-span labels such as `I-USCO`, where recall dropped to **0.01**.

Strengths:

- Data-driven and adaptable
- High performance on frequent patterns
- Easy to retrain on new data

Limitations:

- Lower recall for uncertainty detection and long scopes
- Sensitive to class imbalance
- Requires annotated data and careful feature engineering

8.3. Final Comparison and Outlook

Comparing the two approaches, we observe that the rule-based system performed particularly well in negation detection. With F1-scores around 0.93 for both cue (NEG) and scope (NSCO) identification, it proved to be reliable and consistent, especially in structured clinical contexts. Its strengths include transparency, interpretability, and effectiveness in low-resource settings where labeled data is limited. These qualities make it a good choice for real-world medical applications where high precision and traceability are essential.

However, the rule-based method showed limitations in handling uncertainty detection. The performance for uncertainty cues (UNC) and their scopes (USCO) was notably lower, which suggests that static trigger lists and fixed window scopes are not sufficient to capture the nuanced and variable nature of speculative language. Even after iterative refinement of trigger lexicons, the system struggled with ambiguous expressions, long-distance dependencies, and mixed-language constructs commonly found in clinical texts.

The machine learning model, in contrast, provided a more flexible and data-driven solution. It slightly outperformed the rule-based method in identifying negation cues, reaching an F1-score of 0.95 for B-NEG. While it achieved strong results for frequent and short-span labels, its performance dropped significantly for infrequent and longer spans, particularly for labels like I-USCO. The model often failed to capture the full extent of speculative expressions, highlighting the limitations of using only local lexical and positional features without deeper context. In summary, the rule-based system is better suited for high-precision applications with predictable patterns, while the machine learning approach offers scalability and the potential to adapt to more diverse and complex data. Each method compensates for the other's weaknesses to some extent.

As a future direction, the integration of both systems into a hybrid model may yield better overall performance. Combining the rule-based model's precision with the learning model's adaptability could improve detection accuracy across all label types. Additionally, implementing sequence-aware models, such as Conditional Random Fields (CRFs) or transformer-based architectures like BERT, may enhance scope detection capabilities, particularly for uncertainty, by capturing longer dependencies and richer semantic cues in the text.