Slide 1: Title Slide

Speech:
(Piotrek)
"Hello, we are Group 01 from the Fundamentals of NLP course at UAB. Today, we will present our project on detecting negation and uncertainty in clinical texts using both classical and machine learning techniques. This project is developed by Piotrek, Mimi, Suzana, Adnan, and Iker. We'll take you through the motivation behind the project, the methods we used, the results we obtained, and the future work we're planning."

Slide 2: Introduction

Speech:
(Mimi)
"Our task is to detect negation and uncertainty in clinical narratives. In healthcare, accurately identifying these expressions is critical. For example, when a clinical note says 'no evidence of pneumonia', it implies a definitive conclusion about the absence of the condition. On the other hand, 'possible infection' indicates uncertainty, and further tests may be needed. These expressions affect how clinical data is interpreted, which directly impacts decision-making, coding, and even research outcomes.

We focus on detecting these cues—whether they represent negation or uncertainty—and identifying their respective scopes. The text affected by these cues is crucial to understanding the overall clinical context. To achieve this, we use labels such as NEG for negation cue, NSCO for negation scope, UNC for uncertainty cue, and USCO for uncertainty scope."

Slide 3: Data Exploration and Challenges

Speech:
(Suzana)
"Our dataset consists of clinical notes written in both Spanish and Catalan. These texts are often unstructured, making it difficult to process. They contain a mix of domain-specific terms and abbreviations, which vary from hospital to hospital and even from doctor to doctor. Additionally, these clinical documents often follow a telegraphic style, where sentences are incomplete or lack the standard grammar used in formal writing.

One of the biggest challenges in working with this data is noisy formatting. Clinical texts often have inconsistent punctuation, redacted patient information, and fragmented phrases. This makes tokenization—the process of breaking text into words or meaningful components—tricky. Our dataset also contains both short and long passages, with some containing only 2 cues, while others contain as many as 10 cues, which makes it challenging to build a system that can handle both cases effectively.”

Slide 4: Rule-Based Approach – Overview

Speech:
(Iker)
“To address this problem, we started with a **rule-based approach**, which is both interpretable and simple. We drew inspiration from the **NegEx system**, which is widely used in detecting negation in clinical text. The idea is to use predefined lexicons to identify cues for both negation and uncertainty. These cues are then mapped to a fixed token window of **±5 tokens** around each cue to define the scope of influence.

For example, in a sentence like ‘no hay evidencia de neumonía’ (no evidence of pneumonia), the cue ‘no hay evidencia de’ (no evidence of) is a negation, and we would define the scope to be ‘neumonía’ (pneumonia).

We also applied **preprocessing** to the text to normalize it, which includes lowercasing, accent removal, and tokenization. This ensures that we can consistently process different types of documents and medical jargon.”

Slide 5: Rule-Based Implementation Details

Speech:
(Adnan)
“Building this rule-based system involved creating two curated lists: one for **negation triggers** and another for **uncertainty triggers**. For negation, we included expressions like ‘no hay evidencia de’ (no evidence of), ‘niega’ (denies), and ‘sin’ (without). For uncertainty, we used words such as ‘posible’ (possible), ‘probable’ (probable), and ‘sugiere’ (suggests).

We applied **preprocessing** steps like lowercasing and removing accents, ensuring that the model doesn’t misinterpret words with different cases or characters. When we detected a cue, we applied a window of **5 tokens before and after** the cue to define the **scope**. We also mapped the annotations to the **character level** to facilitate precise evaluation.”

Slide 6: Rule-Based System Evaluation

Speech:
(Piotrek)

"Our rule-based system performed quite well when it came to **negation detection**. The precision and recall for **NEG** and **NSCO** were high, reflecting its accuracy in detecting the cues and their immediate scope. However, it struggled with **uncertainty detection**. For example, the system had a **lower recall** for **UNC** and **USCO** because it missed more subtle expressions of uncertainty and multi-token spans.

This is the limitation of rule-based systems: they rely on predefined lexicons, which means they miss expressions that aren't in the lexicon. Additionally, rule-based systems don't have the ability to understand context as deeply as machine learning models."

Slide 7: Machine Learning Approach – Overview

Speech:
(Mimi)
"To improve the performance, we turned to **machine learning**. In contrast to the rule-based approach, machine learning allows the model to learn directly from the data without needing predefined rules. We used a **supervised learning approach** with a **BIO tagging scheme** to label tokens in the text. This meant we could automatically classify tokens as **part of a negation cue**, **part of the scope**, or **outside of any cue or scope**.

We trained the model using a **linear classifier**, specifically the **SGDClassifier**. The data was split into **80% training** and **20% testing**, so we could evaluate how well the model generalized to new data."

Slide 8: Machine Learning Details

Speech:
(Iker)
"The machine learning model used several features to make its predictions. We first **preprocessed** the text by tokenizing it based on whitespace and removing punctuation. Then, each token was assigned a **BIO tag**: **'B-'** for the beginning of a cue, **'I-'** for tokens inside the scope, and **'O'** for tokens outside of any cue or scope.

We extracted **lexical features**, such as whether the token was capitalized or contained digits. **Contextual features**, like the neighboring tokens, were also included. Additionally, we used **positional features** to mark whether a token was at the beginning or end of the sentence. These features were vectorized using **DictVectorizer** and encoded with **LabelEncoder** for classification."

Slide 9: Machine Learning Evaluation

Speech:
(Adnan)
"Our machine learning model achieved an **overall accuracy of 95%**. It performed well on **negation cues**, achieving high **precision** and **recall**. However, it faced challenges with **uncertainty cues**. The **recall** for **UNC** and **USCO** was lower because the model had difficulty detecting subtle uncertainty expressions or when the scope of uncertainty extended across multiple tokens.

For example, in sentences like 'posible infección' (possible infection), the model sometimes failed to capture the full scope of uncertainty."

Slide 10: Conclusions and Future Work

Speech:
(Suzana)
"To conclude, the **rule-based system** offered a strong baseline for detecting negation, providing **high precision and recall**. However, its coverage was limited, especially for uncertainty detection. On the other hand, the **machine learning model** proved to be more **flexible** and **adaptable**, but it needs further refinement, especially in handling **multi-token uncertainty cues**.

Looking to the future, we plan to experiment with **sequence models** such as **Conditional Random Fields (CRFs)** or **transformers**, which are better suited for capturing the **context** of the text. Combining these models with the **rule-based system** could result in a more **robust and reliable system** that performs well on both negation and uncertainty detection."

Slide 11: Thank You

Speech:
(Piotrek)
"Thank you very much for your attention. We are happy to answer any questions you might have."