Daniel Vidal, 1634599; Neil de la Fuente, 1630223
Paula Feliu, 1630423; Roger García, 1633372

Leveraging NLP Techniques for Negation and Uncertainty Detection in Medical Texts: A Comparative Study of Rule-based and Deep Learning Approaches

Introduction:

The field of natural language processing (**NLP**) has opened up numerous research avenues due to the increasing curiosity for extracting and comprehending medical text data. With the vast quantity of written medical information available, appropriate interpretation can hugely advance our capacity to conduct epidemiological studies, streamline clinical decisions and foster medical research.

One of the biggest challenges in analyzing medical text is deciphering language components like uncertainties and negations. These elements play a critical role in the clinical context, where they can drastically change the meaning of medical statements. To illustrate, the statements "Possible signs of infection" and "No signs of pneumonia" may appear similar, but the nuances between them are significant and require accurate interpretation to prevent faulty conclusions.

Our project goes deep into the field of Spanish medical texts, with a key goal of establishing a methodology that could accurately identify negations, uncertainties, and their respective scopes within texts. We pursued two distinct models to succeed with our goal: a rule-based approach and a Deep Learning model, **Named Entity Recognition** (NER) with BERT (Bidirectional Encoder Representations). Each of these approaches offered different results and performances.

Data cleaning and Management:

We received a JSON file with all the data about the medical text and the annotations, but this data was stored with a structure a little bit confusing and a little hard to extract so we create a DataFrame reading the JSON file one time to get the data in a table like this:

ID	Text	Labels
128391284	n historial clinic del paciente **** fecha de naixement **** ….	[(1332, 1339, **NEG**), (1340, 1348, **NSCO**), (3123, 3129, **UNC**), (3130, 3141, **USCO**)...]
129537734	n historial clinic del paciente **** fecha de naixement **** ….	[(1831, 1838, **NEG**), (1840, 1849, **NSCO**), (3123, 3129, **UNC**), (3130, 3141, **USCO**)...]
…	…	…
133857693	n historial clinic del paciente **** fecha de naixement **** ….	[(1332, 1339, **NEG**), (1340, 1348, **NSCO**), (3123, 3129, **UNC**), (3130, 3141, **USCO**)...]

In this way, we have the labels for every text and all texts and labels in a single data structure. Here the ‘labels’ column consists of a list of tuples that have the indices of start and end in the text and the corresponding label for the words between those indices in the text.

Initially, we considered normalizing the text for training the deep learning model. However, this approach proved ineffective because changing the texts caused the indices to no longer correspond to the words associated with the given labels. Nonetheless, we achieved good results by training the model using the original unnormalized texts.

RULE-BASED METHOD

The rule-based approach for detecting negations, uncertainties, and their scopes begins by creating a ground truth for evaluation. Two lists are created, one for negations and the other for uncertainties, each containing tuples with the scope in words and the corresponding start and end indices. The same process is applied to both negations and uncertainties.

The text is processed using linguistic annotations provided by the "es_core_news_sm" language model, specifically trained for the Spanish language. The next step involves extracting the ground truth negations and uncertainties and adding them to their respective lists. These words serve as rules for detecting negations and uncertainties in future texts.

To find the negations and their scopes, a function called 'find_negations_and_scope' is defined. This function searches for negation words in a list of tokens and identifies the ranges (scopes) where the negations occur. This function is then applied within another function that traverses through all the documents to find the negations and their scopes. The same process is repeated for uncertainties using the 'find_uncertain_and_scope' function.

After obtaining all the predictions from the previous steps, the model is evaluated using two metrics: precision and recall. For precision, two functions are defined. One function compares the words classified as negations or uncertainties with the ground truth, calculating the number of correct predictions and dividing it by the total number of predicted tokens to determine precision. The other function evaluates the precision for scopes by checking if the start and end indices of predicted scopes fall within 10 units of the true scopes.

The recall is calculated using similar procedures. Two functions are created to calculate the recall for identified words and scopes. The function for words compares the list of true values with the list of predicted tokens, counting the number of correct predictions and calculating recall as the proportion of correct predictions out of the total number of true values. The function for scopes compares the list of true scopes with the list of predicted scopes, counting the number of correct scope predictions based on the start and end indices, and calculating recall as the proportion of correct scope predictions out of the total true scopes.

Daniel Vidal, 1634599; Neil de la Fuente, 1630223
Paula Feliu, 1630423; Roger García, 1633372

RESULTS:

From the results, it is clear that the model performs differently across various categories. Let's look at each of them:

1. **Negation**: The model has a high precision of 0.893 which means that when it predicts a negation, it is very likely to be correct. However, the recall score is only 0.232, implying that it misses a large proportion of actual negations in the dataset. This results in a low F1 score of 0.369, which indicates that the balance between precision and recall is not ideal.
2. **Negation Scopes**: Precision drops to 0.450 and recall is 0.349. This means the model is less reliable when predicting the exact scope of negations. While it still correctly identifies a fair number of them, it is also missing a considerable amount and falsely predicting others. The F1 score of 0.393 reflects this somewhat poor performance in both precision and recall.
3. **Uncertainty**: The model's precision is perfect (1.000), which means that every prediction it makes in this category is correct. However, the recall is 0.484, so the model is not identifying all the instances of uncertainty. The F1 score is 0.652, showing a decent balance between precision and recall, although it is biased towards precision due to the perfect score.
4. **Uncertainty Scopes**: Both precision and recall are low (0.343 and 0.189 respectively), indicating the model struggles with identifying the exact scope of uncertainty. The low F1 score of 0.244 confirms this, as it represents the harmonic mean of precision and recall.

Deep Learning Approach:

For the deep learning approach, we implemented a **Name Entity Recognition (NER)** model using the Spacy library, which utilizes **BERT** for training. To adapt the model to our specific problem, we fine-tuned it using a Spacy model trained on a large Spanish language model. We defined our entities of interest as 'NEG' (negations), 'NSCO' (negation scope), 'UNC' (uncertainty), and 'USCO' (uncertainty scope).

To train the model, we needed to format the data in a specific way. The format involved using doc objects (documents from Spacy) that contain span objects representing the entities. These entities were defined with the labels obtained from the data about the medical texts, where we extracted the indices and labels of the negations and uncertainties along with their scopes. We split the data into training and validation sets and stored them in the train.spacy file and validation.spacy file, respectively, in the aforementioned format.

To train the model, we loaded the train and validation files along with a configuration file that contains the model's parameters and settings, which we adjusted to achieve good results. The model was trained with a batch size of 72, maximum epochs of 12, a dropout rate of 0.1, L2 regularization with weight decay, and a learning rate of 0.001.

Daniel Vidal, 1634599; Neil de la Fuente, 1630223
Paula Feliu, 1630423; Roger García, 1633372

Evaluation of the model:

For evaluating the model we did a **qualitative** and **quantitative** evaluation:

The **qualitative evaluation** consisted of showing some examples of entity predictions using some texts from the validation test and some medical texts from the internet. The results we obtained were quite good, the model was able to correctly recognize the negations and uncertainties for all texts, the only problem was that sometimes the scopes were not exactly the same as in the original data or sometimes the scope is cut because of a word that is considered a negation. We also tried with texts that are not medical, for example legal texts or texts about sport and we saw that the model is also able to generalize the negations and the uncertainties but the scopes were mostly imprecise and sometimes there were scopes that did not belong to anything. We think that this imprecision in the scopes is due to the fact that the model is not used to see contexts that are not related to medicine.

For the **quantitative evaluation** we used the metrics that spacy provides to evaluate the model during the training and we got that the model has f-score = 88.00 precision = 86.77 recall = 89.27.

Conclusions:

In conclusion, we can state that the rule-based model for negation detection in NLP performs reasonably well. The F-score indicates a decent balance between precision and recall, resulting in a high accuracy. However, it is expected that the rule-based model is outperformed by the NER model using BERT.

However, the NER model using BERT demonstrates superior performance compared to the rule-based model. It achieves higher accuracy and recall in identifying negations and determining their scope in the text. The NER model captures a larger percentage of negation instances and their scope, resulting in more accurate predictions.

Both models have shown good performance not only on the tested medical texts but also on texts unrelated to medicine. The deep learning model demonstrates the ability to generalize negations and uncertainties across different domains as legal texts or sports-related texts. However, it may encounter difficulties in accurately detecting scopes in non-medical contexts so it might be needed a fine tuning technique to improve its performance.

In conclusion, both the rule-based and NER models have achieved their initial objectives for the specific task they were designed for. The NER model, leveraging BERT, offers improved accuracy, recall, and generalization capabilities compared to the rule-based approach.