George Vlad Manolache 1718986
Darian Balagiu 1719581
Petru Balan 1719379
Valentin Micu Hontan 1718971
Horia Andrei Moraru 1710314

DELIVERY 1 REPORT

RULE-BASED MODEL

HYPOTHESIS

Our hypothesis is based on the next statements:

1. Negation and uncertainty cues are easier to find than medical terms, so we will start out search with the former
2. NEG and UNC amount to much more constrained sets of words than NSCO and USCO, so we can expect better precision on the former
3. Scopes have unpredictable length and are a lot more likely to extend on various words
4. We need to generate a corpus that contains medical terms that are likely to appear in documents

CORE ALGORITHM

All of the approaches that we have tried share the same starting steps:

1. Create two sets for Negation and Uncertainty cues called NEG and UNC
2. Parse the documents from the training set. Parse the Ground Truth NEG and UNC labels and add the cues to the two sets created above.
3. The search will be based on the cues: we will create two regexes by putting together all the words from NEG and UNC using “|” operations. We will match the regexes on each text and for each match found, we will then employ a particular algorithm to find the corresponding scope.

NOTE: This approach is exactly the opposite to NegEx, in which they were searching for medical terms and then verified if they are inside a negation/uncertainty cues’ scope.

ANALYZING THE DATA

We start by creating lists of highlighted words (NEG, UNC, USCO, NSCO) for the training set and test set in order to compare similarities between them. These are results:

	NEG	UNC	NSCO	USCO
WORDS FROM TEST ALSO IN TRAINING	1118/1132	118/131	599/1074	20/129

Table 1

Judging by this table, we can expect high precision for both of the cues, medium precision for NCSO and bad for UCSO. That is because, as mentioned in the hypothesis, cues are part of a much more constrained set of words than scopes.

There are some words from NEG that are also found in UNC (such as some occasional “no”, “not”). Keeping these words in UNC will cause a lot of noise because the GT rarely categorizes them as UNC instead of NEG. That is why we are going to remove words from UNC that are also in NEG.

APPROACH 1: CUTEXT Medical Terminology Parser

CUTEXT is a tool used for extracting medical terms from a corpus. Because of the lack of terminology in the medical field (and also our inability to read Catalan to extract useful information from the documents) we tried creating our own medical corpus. The resulting medical terms generated a set called CUTEXT_SCOPES, then a regex was created by piping together (“|”) all the terms from it, and for each cue found, the regex was being matched to the next or previous 5 words. This was successful, however in the end the results were not great:

	NEG	NCSO	UNC	USCO
PRECISION	92.3%	15%	56%	8%
RECALL	98%	17%	77%	11%
F1	95%	16%	65%	9%

The reason for that is because there was not enough similarity between our personally created corpus and the test set.

APPROACH 2: BASELINE

Our BASELINE approach starts by following the core algorithm mentioned above to create the NEG and UNC sets for cue prediction and uses the same idea to create NCSO and UCSO sets.

If we take a look at Table 1, we can expect this approach to yield an average performance on NCSO predictions of around 56% (599/1074).

It is worth noting that both the CUTEXT and Baseline approach work with the same NEG and UNC cues.

1. Create a set called ALL_SCOPES that combines NCSO and UCSO (our search is going to be done for the cues, so there is no reason to separate the scopes) and a regex pattern that pipes all of the scopes together
2. Whenever we match a cue, use the scope regex pattern above on the previous 5 words or next 5 words

   We are aware that 44% of NCSO scopes and 85% of UCSO will not be found because they were not learnt from the training set, but this is just a Baseline approach. We will try to find optimizations to solve this problem later.

After running the BASELINE model on the training set and test set provided, we obtained the following results:

	NEG	NCSO	UNC	USCO
PRECISION	91.1%	43.3%	53.3%	8%
RECALL	97%	48.5%	73.3%	12%
F1	94%	45.7%	61.7%	10%

The NEG prediction performs very well and rises to the level of expectation.

The NCSO prediction does not reach its maximum potential of this approach (56%) and that is because we only took into consideration that the scope will be after the cue. This is a viable optimization that we can undertake.

The UNC does no rise to the expectations and that is because of the words that we removed some useful words from the set (because they were also in NEG and hence the model would not be able to distinguish between the two, especially if it is only rule-based)

APPROACH 3: BASELINE + CUTEXT

This was done by combining the NCSO and UCSO sets of the previous two methods. The following results are obtained:

	NEG	NCSO	UNC	USCO
PRECISION	92.5%	51.7%	57%	21%
RECALL	97%	57%	77%	28.7%
F1	94.7%	54.2%	66%	24%

Improvements can be seen in all areas.

APPROACH 4: SCOPES ARE EXTENDED UNTIL END OF SENTENCE

This approach is probably the simplest one and yet produces the best results for NCSO and UCSO, but the worst ones for NEG and UNC.

	NEG	NCSO	UNC	USCO
PRECISION	94%	61%	56%	28%
RECALL	84%	58%	68%	35%
F1	89%	60%	61%	31%

DELIVERY 2 REPORT

MACHINE LEARNING

1. INTRODUCTION

This paper describes our approach to solving the task of automatically identifying all the negation and uncertainty cues in a medical document along with their scopes, including discontinuous negative expressions which, as we will show in the following sections, is the most challenging aspect in this task.

In this report we are presenting our approach to this problem by using a CRF model with various features that will be detailed in the following sections.

This paper is organized as follows: Section 2 presents the data set that we are going to work with. In Section 3 the techniques that were used for preprocessing and the solutions to the irregularities found in the data set are presented. Section 4 showcases the algorithms used for training the CRF. The results achieved by our approach are described and briefly analyzed in Section 5.

2. REVIEWING THE DATA SET

As CRF is a **supervised machine learning model**, for solving this task we are using a training corpus consisting of 254 documents containing medical text reviews on people. Each document contains Ground Truth results for each of the 4 assignable categories along with the corresponding positions in the text at which we can find those words.

As can be seen in the following example extracted from the first document from the training set, the word that starts at position 449 and ends at 452 belongs to the “NEG” category can be found in the text as “no”, which clearly signifies a negation cue.


```
{
        "start": 449,
        "end": 452,
        "labels": ["NEG"]
}
```


```
per **************, *****; **************, ****** informe d'alta d'hospitalitza
cloramfenicol . no habitos toxicos. antecedentes medicos: bloqueo auriculoventricu
lesiones cutaneas con anestesia local protesis total de cadera cordectomia hernior
```

The next item in the Ground Truth is represented by the scope of the previous negation cue, labeled as “NSCO”. It starts at 452 and ends at 468 and is represented by “habitos toxicos”.


```
{
        "start": 452,
        "end": 468,
        "labels": [ "NSCO"]
}
```



```
"text": " nº historia clínica: ** *** *** nºepisodi: ******** sexe: home data de naixement: 16.05.19.
per ***************, *****; ****************, ****** informe d'alta d'hospitalitzacio motiu d'ingres pacie
cloramfenicol . no habitos toxicos. antecedentes medicos: bloqueo auriculoventricular de primer grado hipe
lesiones cutaneas con anestesia local protesis total de cadera cordectomia herniorrafia inguinal proces ac
```


3. PREPROCESSING

The first step to solving this task is to preprocess our data so that we can extract useful features from it and train our CRF model with them. When analyzing the texts, we came across various difficulties to which we are going to explain their solutions below:

3.1. LANGUAGE MIXTURE

The first issue that we tackled is that this data set is using both Catalan and Spanish. This becomes a problem when trying to modify and extract information from the text (by lemmatizing, POS tagging etc) because the language used is inconsistent.

Our solution to this problem was to use the **spacy** framework’s ca_core_news_md and es_core_news_md methods that are meant to preprocess and tag each sentence of a text for Catalan, respectively Spanish. However, we found out that only applying the function for Spanish yields better results, so we stuck only with Spanish. The reason for this may be either that the text is composed mostly by Spanish words or that the Catalan version of this library is not as effective as the Spanish one because of its lower utilization.

3.2. ANOMALIES IN THE GROUND TRUTH

There are some particular words in the document that are being tagged in a mysterious matter. For example, “exfumador” and other similar words are being tagged as a NEG and NSCO at the same time.

```
 servei otorrinolaringologia data d'ingres 18.06.2018 data d'alta 19.06.2018
 alergias medicamentosas conocidas exfumador de 8 años 1trasplante bipulmona
con 28%pmn. ac anti-hla negativos. 3.- diabetes mellitus tipo 2 : buenos
```

In the 18th document we can see one such problem in which the first two letters of the word (“ex”) are categorized as NEG cue

```
{
            "start": 397,
            "end": 399,
            "labels": [
              "NEG"]
}
```

while the rest of the word is categorized as NSCO.

```
{
            "start": 399,
            "end": 406,
            "labels": [
              "NSCO"
            ]
          }
```


Splitting a single word in two different tags is inconsistent in the corpus and is a source of errors. We avoided this problem by treating this particular case before applying the general rules for the preprocessing part. Therefore, we managed to alleviate this problem by doing **data imputation**.

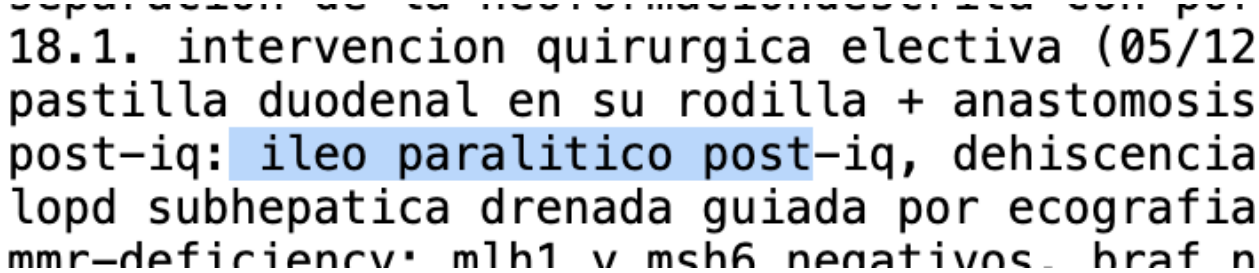
```
#move to the next word if didn't find the beginning of a negation
lst = ['exfumador','exfumadora','ex-fumador','ex-fumadora']
if words_info[words_counter]['word'] in lst:
  words_info[words_counter]['tag'] = 'B-NEG'
  neg_counter+=1
  words_counter+=1
  continue
```

3.3 INCONSISTENT TAGGING IN GROUND TRUTH

Through thorough analysis of the training set documents we found some cases of inconsistent tagging. Some such cases are those for which tagging of the scope ends in the middle of the word, especially for words that are separated by a line.

In this following example, the negation scope is ending right before the “-” and since our parser is only separating words by spaces or end of sentence, this inconsistent tagging is generating parsing errors.

Solving this problem would mean changing our whole parsing technique, but luckily these types of errors only occur in 7% of the documents, so we chose to not use those documents at all in training.



Therefore, after reviewing each document from the training set, we ended up with 234 training documents out of the initial 254.

4. TRAINING THE CRF

4.1 TAGGING THE WORDS

The first step in training the CRF is tagging the words from the training data. The tagging method we used is **BIOE Tagging,** consisting of a total of 13 tags.

	Total number of appearances	Percentage of tag out of total words	Observations
TOTAL WORDS IN 234 DOCS	163485	100%	
O	145826	89.19%	Labeled as Other
B-NEG	3877	2.37%	Labeled as beginning of negation cue
B-NSCO	3713	2.27%	Labeled as beginning of negation scope
B-UNC	418	0.26%	Labeled as beginning of uncertainty cue
B-USCO	411	0.25%	Labeled as beginning of uncertainty scope
I-NEG	0	0%	Labeled as inside of negation cue
I-NSCO	5253	3.21%	Labeled as inside of negation scope

I-UNC	4	0.0025%	Labeled as inside of uncertainty cue
I-USCO	879	0.53%	Labeled as inside of uncertainty scope
E-NEG	76	0.04%	Labeled as ending of negation cue
E-NSCO	2511	1.53%	Labeled as ending of negation scope
E-UNC	198	0.12%	Labeled as ending of uncertainty cue
E-USCO	319	0.2%	Labeled as ending of uncertainty scope

Upon tagging all the sentences, the following statistic were obtained:

Number of Sentences	Number of Sentences that contain only 'O' tags	Number of Sentences that contain BIE tags
11.150	7573	3577

4.2 FEATURES LEARNT BY CRF

For this task, we were inspired by the CRF created by **Universitat Politecnica de Catalunya** in the paper named "***Negation Cues Detection Using CRF on Spanish Product Review Texts***".

These are the features that we used:

- Bias
- Word in lowercase
- Part of speech
- Lemma
- Suffix
- Word is a digit
- BOS (word is placed at the beginning of a sentence)
- EOS (word is placed at the end of a sentence)
- The part of speech, lemma, isDigit of all the words before and after the current word that are in the range of 3

Here is an example of the features for a word that is placed at the beginning of the sentence.

```
{'bias': 1.0, 'word.lower()': 'no', 'pos': 'ADV', 'lemma': 'no', 'suffix':
'no', 'word.isdigit()': False, '+1:word.lower()': 'habitos', '+1:pos':
'NOUN', '+1:lemma': 'habito', '+1:suffix': 'tos', '+1:word.isdigit()':
False, '+2:word.lower()': 'toxicos', '+2:pos': 'ADJ', '+2:lemma':
'toxico', '+2:suffix': 'cos', '+2:word.isdigit()': False, 'bos': True,
'eos': False}
```

4.3 PARAMETERS OF THE CRF

```
trainer_crf.set_params({
   'c1': 1.0,   # Coefficient for L1 regularization
   'c2': 1e-3,  # Coefficient for L2 regularization
   'max_iterations': 100,
   'feature.possible_transitions': True
})
```

5. MAKING PREDICTIONS

We started by downloading the test set and applying all the preprocessing techniques applied for the training set above. We ended up with 61 documents composed by 3059 sentences.

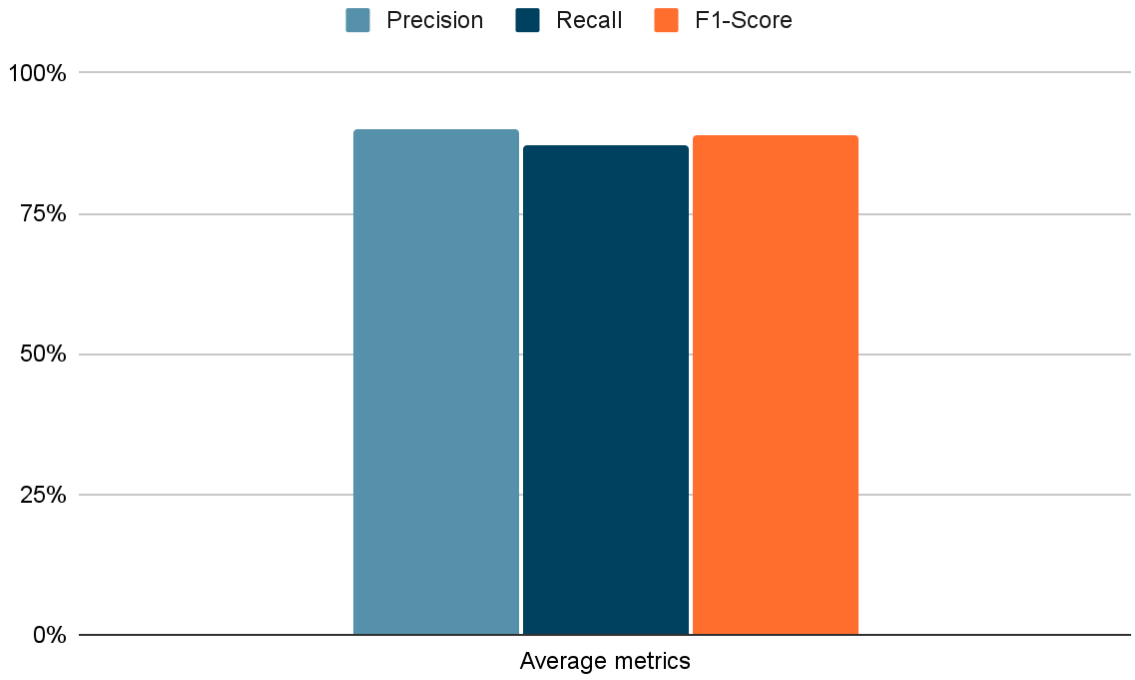
Number of Sentences	Number of Sentences that contain only ‘O’ tags	Number of Sentences that contain ‘BIE’ tags
3059	2088	971

The training set : test set rapport is 79% : 21%, which is a solid repartition of the data.

Then, we applied our CRF model on all of the test set sentences and obtained the following results:

	precision	recall	f1-score	support
B-NEG	0.97	0.96	0.97	1019
E-NEG	0.94	0.89	0.91	18
B-NSCO	0.95	0.92	0.94	971
E-NSCO	0.86	0.86	0.86	643
I-NSCO	0.88	0.88	0.88	1304
B-UNC	0.91	0.71	0.80	116
E-UNC	0.87	0.72	0.79	54
I-UNC	0.00	0.00	0.00	3
B-USCO	0.91	0.72	0.81	115
E-USCO	0.61	0.49	0.54	88
I-USCO	0.69	0.64	0.66	251
micro avg	0.90	0.87	0.89	4582
macro avg	0.78	0.71	0.74	4582
weighted avg	0.90	0.87	0.89	4582
samples avg	0.10	0.10	0.10	4582

It seems that the BIOE tagging is efficient. The only tags that the model struggles with are I-UNC, E-USCO and I-USCO but those categories provided inferior results also in the Rule Based Model. That is happening due to the fact that UNC and USCO words are more ambiguous, but also because they are more sparse and the model doesn't have enough training samples to learn from.



Now, in order to find the exact number of correct predictions we need to identify how many structures start and end at the same word.

The results are the following:

	NEG	NCSO	UNC	USCO
PRECISION	97%	79%	89.1%	46.6%
RECALL	95.7%	84.9%	68.9%	47.4%
F1	96.5%	81.8%	77.7%	47%

6. GROUND TRUTH VS PREDICTIONS

Negation cues
Negation scope
Uncertainty cues
Uncertainty scope

True:

no se palpan masas ni megalias, ni tampoco globo vesical. no signos de irritacion peritoneal. puñopercusion lumbar bilateral negativa. neurologico: vigil y orientada en las tres esferas, no alteraciones del lenguaje, no signos de focalidad neurologica aparentes. exploracio complementaria ecg : taquicardia sinusal a 130lpm. pr 120 mseg, qrs 80mseg. eje 0 sin alteraciones agudas en la repolarizacion. rx torax : indice cardiotoracico no aumentado (<0.5). senos costofrenicos libres. no condensaciones parenquimatosas.

Predicted:

no se palpan masas ni megalias, ni tampoco globo vesical. no signos de irritacion peritoneal. puñopercusion lumbar bilateral negativa. neurologico: vigil y orientada en las tres esferas, no alteraciones del lenguaje, no signos de focalidad neurologica aparentes. exploracio complementaria ecg : taquicardia sinusal a 130lpm. pr 120 mseg, qrs 80mseg. eje 0 sin alteraciones agudas en la repolarizacion. rx torax : indice cardiotoracico no aumentado (<0.5). senos costofrenicos libres. no condensaciones parenquimatosas.

True:

proces actual paciente acude con familiares (a quienes se realiza entrevista indirecta ya que el paciente se encuentra confuso). refieren que hace aproximadamente 48 horas inicia disnea de pequeños esfuerzos y tos sin expectoracion junto con sensacion distermica. el dia de hoy se lo encuentran en el suelo, desorientado pero sin perdida de consciencia, con auscultacion de secreciones abundantes y malestar general, por lo que avisan al sem. aparentemente no ha presentado fiebre termometrada, dolor toracico opresivo, palpitaciones o cortejo vegetativo.

Predicted:

proces actual paciente acude con familiares (a quienes se realiza entrevista indirecta ya que el paciente se encuentra confuso). refieren que hace aproximadamente 48 horas inicia disnea de pequeños esfuerzos y tos sin expectoracion junto con sensacion distermica. el dia de hoy se lo encuentran en el suelo, desorientado pero sin perdida de consciencia, con auscultacion de secreciones abundantes y malestar general, por lo que avisan al sem. aparentemente no ha presentado fiebre termometrada, dolor toracico opresivo, palpitaciones o cortejo vegetativo.

True:

tc: litiasis de 6-7mm a nivel del ureter distal. tumoracion probablemente solida en polo superior del riñon izquierdo de 2,5cm a completar estudio. en tc de contraste de control se evidencian imagenes sugestivas de quistes renales, no neoplasicas; cateter 2j derecho normoposicionado y litiasis de 5mm en ureter pelvico ipsilateral. portador de cateter doble j derecho desde entonces.

Predicted:

tc: litiasis de 6-7mm a nivel del ureter distal. tumoracion probablemente solida en polo superior del riñon izquierdo de 2,5cm a completar estudio. en tc de contraste de control se evidencian imagenes sugestivas de quistes renales, no neoplasicas; cateter 2j derecho normoposicionado y litiasis de 5mm en ureter pelvico ipsilateral. portador de cateter doble j derecho desde entonces.

7. CONCLUSIONS

In this report we have described our approach for the detection of negations and uncertainties cues and scopes based on a CRF classifier. The results obtained show that this supervised learning technique is a promising approach to building a system that automatically detects negation and uncertainties correctly tagging 93% of unseen sentences from the test set (taking into account the ones that don't contain negations or uncertainties) and with an average F1 score of 75.25% on unseen data (negation cues, negation scopes, uncertainty cues and uncertainty scopes. It can be clearly seen that this supervised approach performs far better than the rule based method technique we previously implemented.

In terms of limitations, we observed that the model is not performing well in situations where the cue follows the scope (because of the few training examples that are found in the training dataset). We also think that in some cases the classifier is not able to understand the semantic relationship between the words of the sentence.

DELIVERY 3 REPORT

DEEP LEARNING

1. INTRODUCTION

This paper describes the negation / uncertainty cue detection system in medical documents along with their scopes. The proposed system consists of a deep learning architecture based on the use of a Bi-LSTM to process contextual information.

This work is organized as follows:

1.2 Deep Learning model

The proposed architecture for negation and uncertainty detection employs a multi-modal deep learning approach(Figure 1). It starts with an **embedding layer** for character-level input (Xch), which is processed through a **convolutional layer** followed by **max pooling** and **flattening**. This character-level representation is then concatenated with word-level features (Xw), PoS-tagging features (Xp), and casing features (Xc). The word and PoS-tagging features are processed through their respective embedding layers, while the casing features are one-hot encoded.
We selected as casing features the following:

1) word contains numbers
2) punctuation sign are presented before or after the word
3) is the first word in the sentence
4) is the last word in the sentence

The concatenated features are fed into a **Bidirectional LSTM (Long Short-Term Memory)** network, capturing both forward and backward dependencies. The output from the LSTM is passed through a **dense neural network** which ultimately predicts the target labels (T1, T2, ..., Tz).

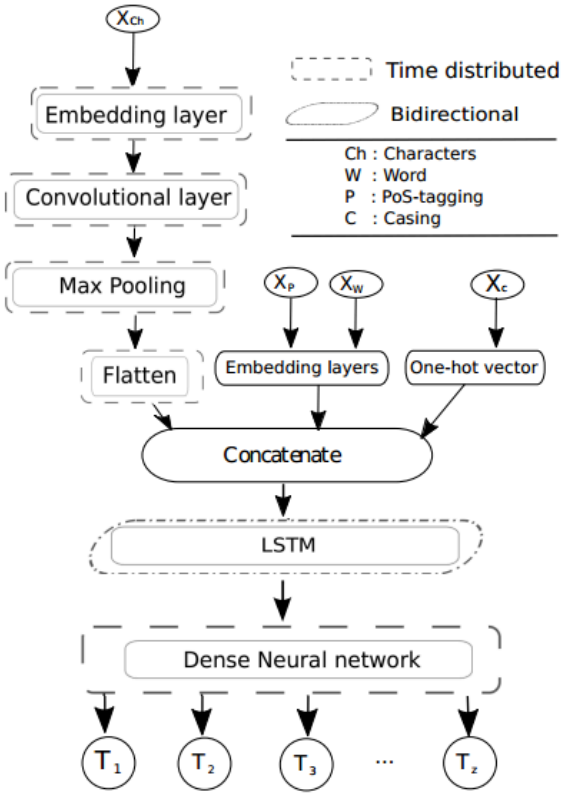


Figure 1

Using the training set, the model has been trained during a total of 20 epochs. Architecture of the model is the following:

– Word Embedding dimension: 300
– Embeddings dimension (Characters / PoS-tagging): 50 / 50

– One-hot dimension (Casing): 8
– Bi-LSTM output dimension: 128 (per each LSTM)
– 2 Dense layers after Bi-LSTM layer: 128 respectively 16 (tag size) for predictions
– Conv2D (kernel size / filter): 3 / 28
– Batch size / Model optimizer: 64 / Adam optimizer [12]

– Dense layers after embedding layers (to adjust the shape of the parameters)

In order to avoid overfitting we used dropout as following:
– Conv2D output dropout: 0.5
– Dense layer dropout: 0.4

Sentence Tagging and processing

The ground truth that the model tries to predict is the sentence tagged with the followings tags that we later index:

B-NEG: Beginning of Negation
E-NEG: End of Negation
B-NSCO: Beginning of Negation Scope
I-NSCO: Inside of Negation Scope
E-NSCO: End of Negation Scope
B-UNC: Beginning of Uncertainty
E-UNC: End of Uncertainty
I-UNC: Inside of Uncertainty
B-USCO: Beginning of Uncertainty Scope
I-USCO: Beginning of Uncertainty Scope
E-USCO: End of Uncertainty Scope
O: Other

Before giving it to the training set as ground truth, we marked the start and end of every sentence with indexes and added padding to give every sentence the same length to prepare it for the data loader.

Models comparison

Model	Rule-Based			Machine Learning			Deep Learning		
Class	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
NEG	92.5%	97%	94.7%	97%	95.7%	96.5%	95.5%	83.5%	88.5%
NCSO	51.7%	57%	54.2%	79%	84.9%	81.8%	76%	78%	77%
UNC	57%	77%	66%	68.9%	68.9%	77.7%	89%	77%	82.5%
USCO	21%	28.7%	24%	46.6%	47.4%	47%	60%	47%	52%

The table presents the performance metrics—precision, recall, and F1 score—of three models (Rule-Based, Machine Learning, and Deep Learning) across four classes: NEG, NCSO, UNC, and USCO. This summary provides an overall comparison of the models' effectiveness in detecting negation and uncertainty in text.

Overall Performance

The Machine Learning model consistently outperforms the other two models across most metrics and classes, demonstrating its superiority at identifying negation cues and scopes. The Deep Learning model also shows strong performance, particularly in comparison to the Rule-Based model.

We think that the Deep Learning model didn’t achieve the best results in all categories because we did not have enough data to train the model.

NEG Class

The NEG class, representing negations, is critical for accurate text analysis.

- **Deep Learning** achieves high precision (95.5%) and a good recall (83.5%), leading to an F1 score of 88.5%. This indicates that the Deep Learning model is highly accurate and reliable in detecting negations.

- **Machine Learning** also performs well, with a precision of 97% and the highest recall (95.7%) among all models, resulting in an F1 score of 96.5%. This suggests excellent performance.
- **Rule-Based** methods show good precision (92.5%) and recall (97%), but their F1 score (94.7%) is lower than that of the Machine Learning model, indicating lower overall effectiveness.

NCSO Class

For the NCSO class, the detection of specific cues is necessary.

- **Machine Learning** excels with a precision of 79%, recall of 84.9%, and F1 score of 81.8%, making it the best performer in this class.
- **Deep Learning** follows with slightly lower precision (76%) and recall (78%), resulting in an F1 score of 77%. This shows that while effective, it is slightly less so than Machine Learning in this category.
- **Rule-Based** methods have the lowest metrics, with a precision of 51.7%, recall of 57%, and F1 score of 54.2%.

UNC Class

The UNC class requires detecting uncertainty.

- **Deep Learning** leads with the highest precision (89%) and an F1 score of 82.5%, despite a recall of 77%. This suggests strong performance in identifying uncertainty with fewer false positives.
- **Machine Learning** achieves equal precision and recall (68.9%), leading to an F1 score of 77.7%, not as high as Deep Learning.
- **Rule-Based** methods show the lowest precision (57%) and recall (77%), resulting in an F1 score of 66%, indicating the least effectiveness among the three models.

USCO Class

The USCO class is often challenging due to its complexity.

- **Deep Learning** again shows the best performance with a precision of 60%, recall of 47%, and F1 score of 52%. This indicates it handles the complexity better than the other models.
- **Machine Learning** follows with a precision of 46.6%, recall of 47.4%, and an F1 score of 47%, performing adequately but less effectively than Deep Learning.
- **Rule-Based** methods perform the worst, with precision (21%), recall (28.7%), and F1 score (24%), highlighting significant limitations in handling this class.

Conclusions

This study introduces a deep learning system for detecting negation and uncertainty cues in medical documents, leveraging a Bi-LSTM architecture. The system integrates character-level, word-level, PoS-tagging, and casing features, processed through a convolutional and Bidirectional LSTM network. Training involved 20 epochs with dropout to prevent overfitting.

In comparing the performance of Rule-Based, Machine Learning, and Deep Learning models, Machine Learning consistently outperformed the others across most metrics and classes, demonstrating its effectiveness in identifying negation cues and scopes. The Deep Learning model showed strong performance, particularly in precision, although its results were limited by the dataset size.

- **NEG Class**: Machine Learning achieved the highest F1 score (96.5%), with Deep Learning also performing well (88.5%).
- **NCSO Class**: Machine Learning excelled with an F1 score of 81.8%, while Deep Learning scored 77%.
- **UNC Class**: Deep Learning led with an F1 score of 82.5%, outperforming Machine Learning.
- **USCO Class**: Deep Learning had the best performance with an F1 score of 52%.

In conclusion, while the Machine Learning model is currently the most effective, the Deep Learning model holds significant potential, particularly with more extensive training data. The Rule-Based model, while useful, lags behind in effectiveness compared to the other approaches.

Bibliography

Rule-based Methods

- **Universidad Politécnica de Madrid**
  - An Approach to Detect Negation on Medical Documents in Spanish
  - Integrating Speculation Detection (2021)
- **University of Pittsburgh**
  - A Simple Algorithm for Identifying Negated Findings and Diseases (2002)

Machine Learning Methods

- **Universitat de Barcelona, Spain**
  - Detection of Negation Cues in Spanish: TheCLiC-Neg System (2019)
- **University of Antwerp**
  - A metalearning approach to processing the scope of negation (2009)
- **Universitat Politècnica de Catalunya**
  - Negation Cues Detection Using CRF on Spanish Product Review Texts (2018)
- **University of Oslo, Department of Informatics**
  - An open-source tool for negation detection: a maximum-margin approach (2017)

Deep Learning Methods

- **Universidad Nacional de Educación a Distancia (UNED)**
  - University of Oslo, Department of Informatics (2018)
- **H. Fabregat, A. Duque, L. Araujo**
  - Extending a Deep Learning Approach for Negation Cues Detection in Spanish (2019)
- **University of Edinburgh**
  - Neural Networks For Negation Scope Detection (2016)