July 2023

Enhancing the Quality of Parallel Corpora through Classic and Neural Classification

Ana Rodero Paredes

Table of contents

01
Introduction

02
State of the art

03
Experiments

04
Results

05
Conclusions

01

Introduction

Introduction

Evaluation of translation quality

- BLEU
- Other metrics

Parallel corpora filtering

- Bifixer
- Bicleaner
- OpusFilter
- OpusCleaner

02

State of the art

Parallel corpora filtering

BIFIXER

Remove empty lines

Fixes several text issues

Orthography fixing

Tokenization fixing

Segmentation of long sentences

Removal of duplicates

- Hash
- Ranking

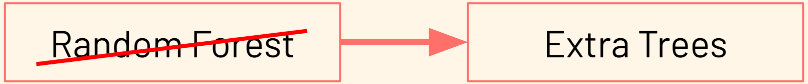
Parallel corpora filtering

BICLEANER

Likelihood of being mutual translations

```
bicleaner-train
```

```
bicleaner-classify
```



~~Random Forest~~ → Extra Trees

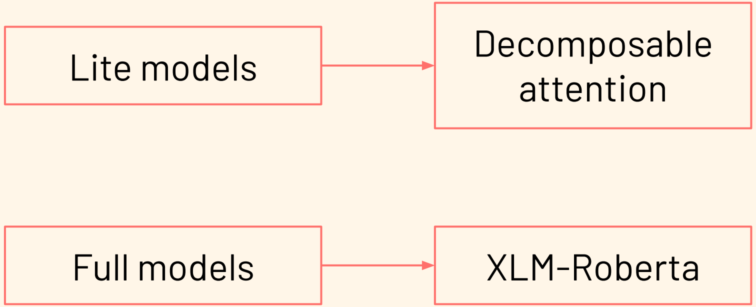
Parallel corpora filtering

BICLEANER AI

Likelihood of being mutual translations

`bicleaner-ai-train`

`bicleaner-ai-classify`

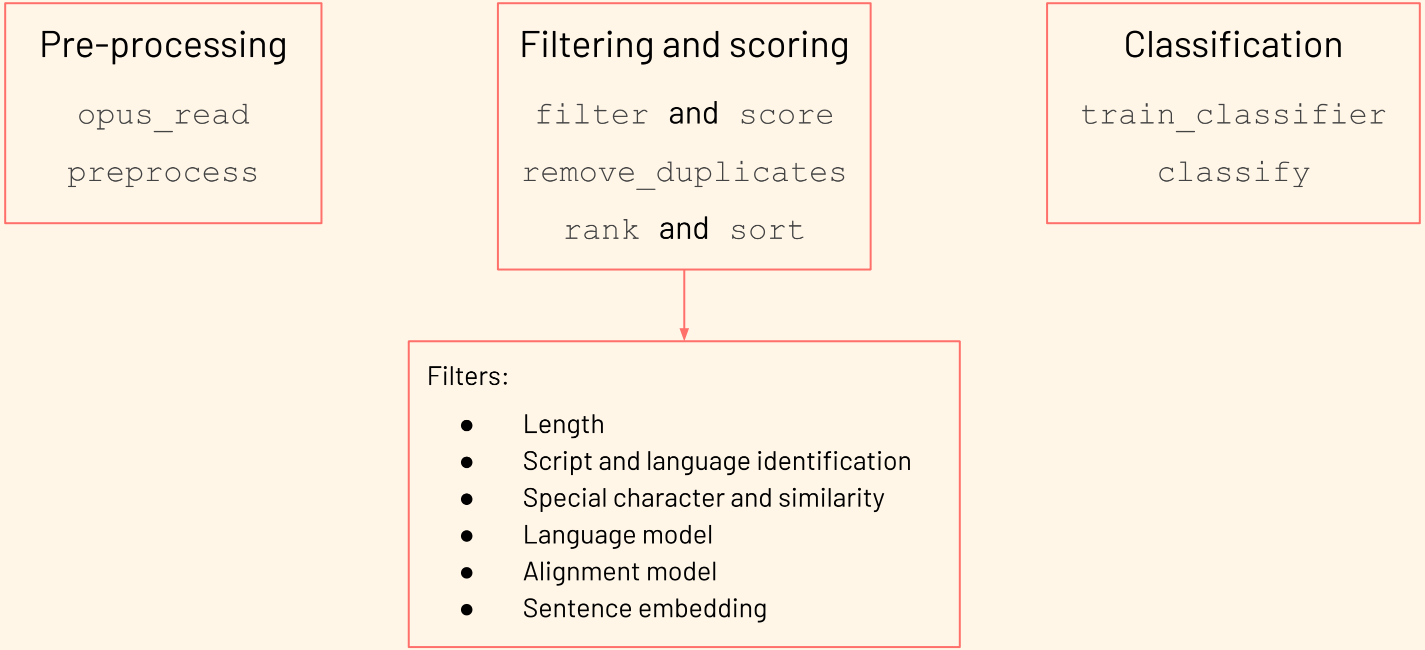


Lite models → Decomposable attention

Full models → XLM-Roberta

Parallel corpora filtering

OPUSFILTER



Pre-processing

`opus_read`

`preprocess`

Filtering and scoring

`filter` and `score`

`remove_duplicates`

`rank` and `sort`

Classification

`train_classifier`

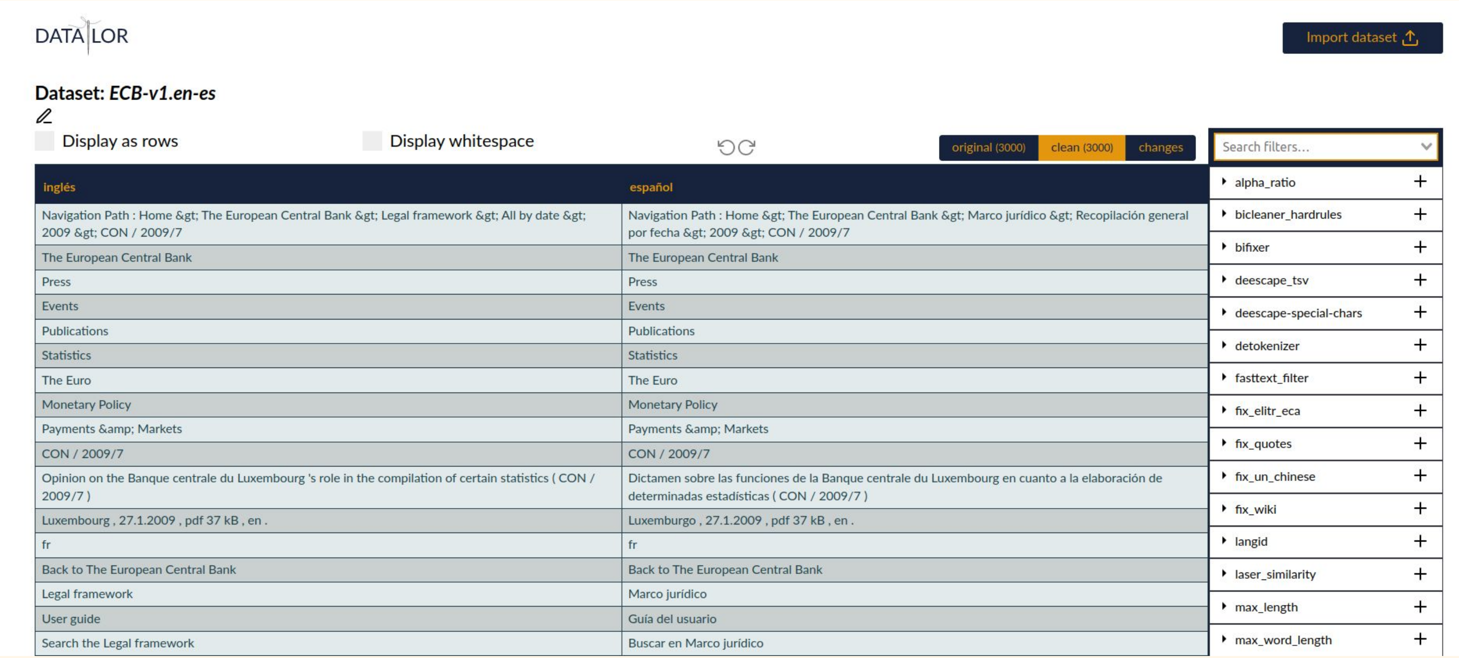
`classify`

Filters:

- Length
- Script and language identification
- Special character and similarity
- Language model
- Alignment model
- Sentence embedding

Parallel corpora filtering

OPUSCLEANER



Dataset: *ECB-v1.en-es*

inglés	español
Navigation Path : Home &gt; The European Central Bank &gt; Legal framework &gt; All by date &gt; 2009 &gt; CON / 2009/7	Navigation Path : Home &gt; The European Central Bank &gt; Marco jurídico &gt; Recopilación general por fecha &gt; 2009 &gt; CON / 2009/7
The European Central Bank	The European Central Bank
Press	Press
Events	Events
Publications	Publications
Statistics	Statistics
The Euro	The Euro
Monetary Policy	Monetary Policy
Payments &amp; Markets	Payments &amp; Markets
CON / 2009/7	CON / 2009/7
Opinion on the Banque centrale du Luxembourg 's role in the compilation of certain statistics (CON / 2009/7)	Dictamen sobre las funciones de la Banque centrale du Luxembourg en cuanto a la elaboración de determinadas estadísticas (CON / 2009/7)
Luxembourg , 27.1.2009 , pdf 37 kB , en .	Luxemburgo , 27.1.2009 , pdf 37 kB , en .
fr	fr
Back to The European Central Bank	Back to The European Central Bank
Legal framework	Marco jurídico
User guide	Guía del usuario
Search the Legal framework	Buscar en Marco jurídico

03

Experiments

Training the base models

Bicleaner classic

Dictionaries: Opensubtitles: 4M + Eubookshop: 2M + Tilde: 2M + DGT: 2M
Training corpus: Newscommentary: 100K

1) Clean data:
    a) Detokenize
    b) Bifixer
    c) Hardrules
2) Dictionaries
3) Word frequency files
4) Train model

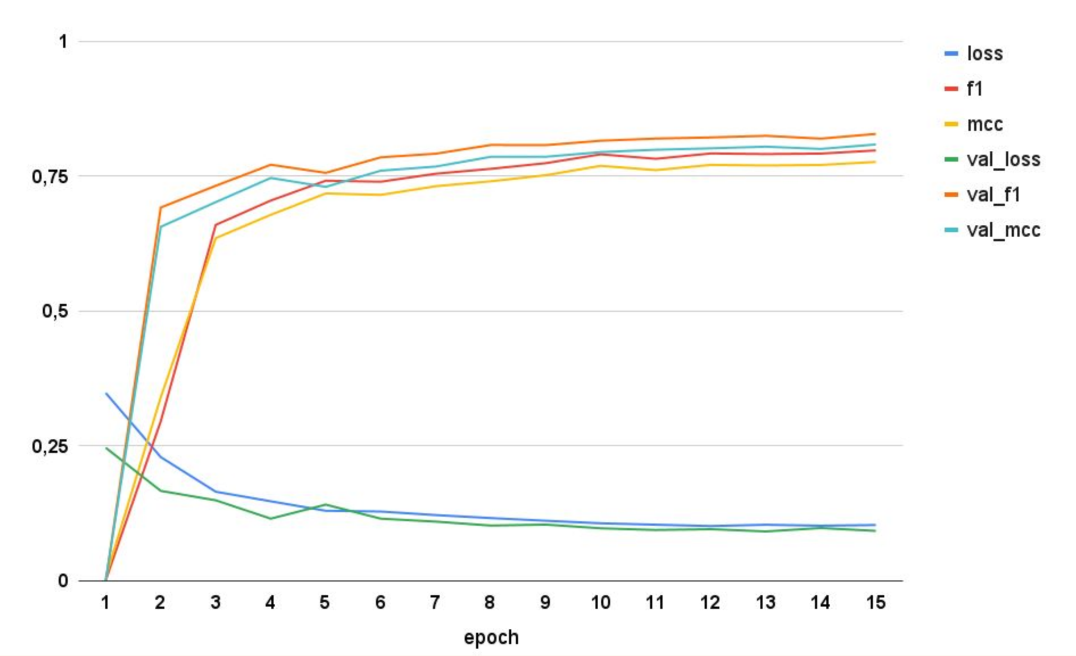
Bicleaner AI full

Corpus: Opensubtitles, Tilde, JW300, Europarl, TedTalks, GlobalVoices, NewsCommentary, UNPC, Scielo

1) Clean data:
    a) Detokenize
    b) Bifixer
    c) Hardrules
2) Word frequency file (10M)
3) Train model (700k / 2k)
    a) Batch size: 16
    b) 15 epochs
    c) 3000 steps/epoch
    d) Patience: 3

Training the base models

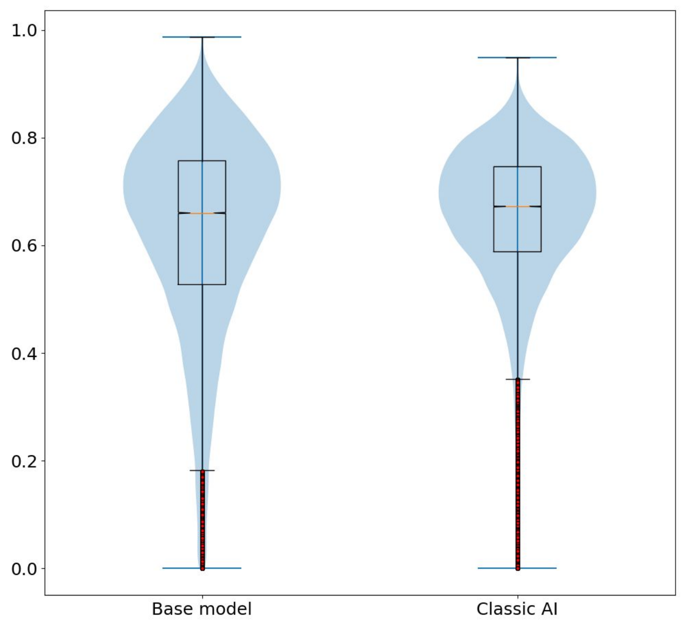
Bicleaner AI full

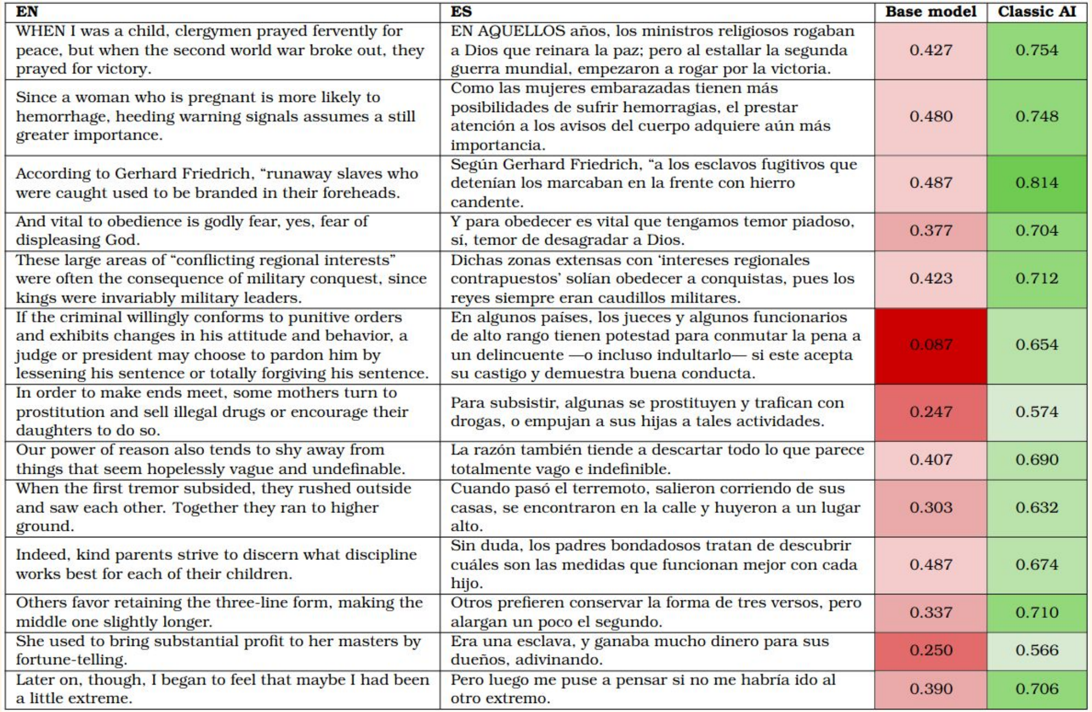


- Precision: 0.841
- Recall: 0.883
- F1: 0.862
- MCC: 0.846

Improving Bicleaner classic

Using AI model data to train classic model



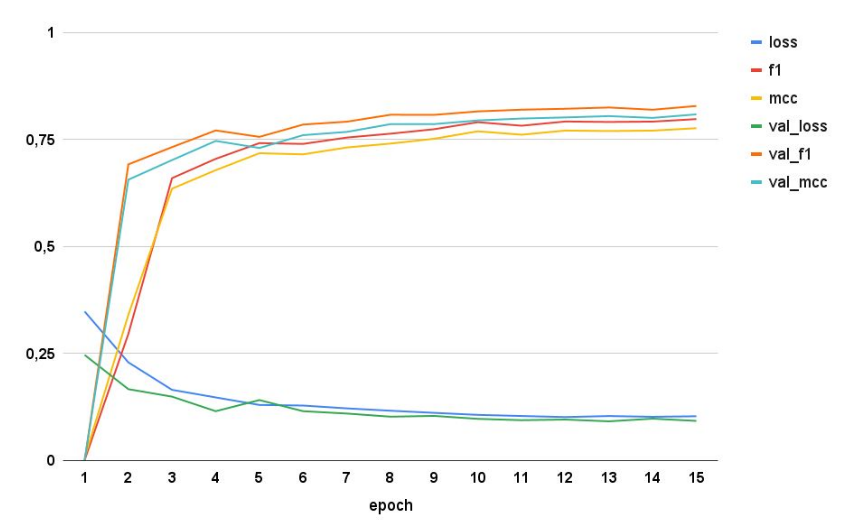


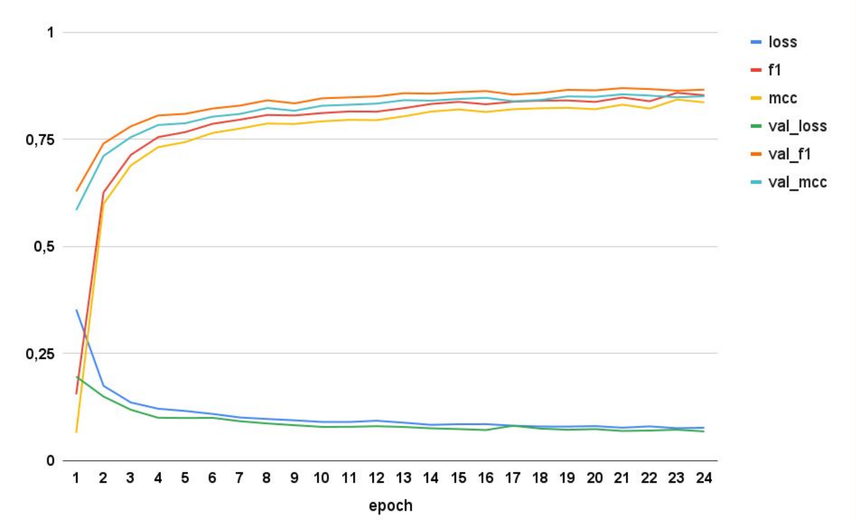
EN	ES	Base model	Classic AI
WHEN I was a child, clergymen prayed fervently for peace, but when the second world war broke out, they prayed for victory.	EN AQUELLOS años, los ministros religiosos rogaban a Dios que reinara la paz; pero al estallar la segunda guerra mundial, empezaron a rogar por la victoria.	0.427	0.754
Since a woman who is pregnant is more likely to hemorrhage, heeding warning signals assumes a still greater importance.	Como las mujeres embarazadas tienen más posibilidades de sufrir hemorragias, el prestar atención a los avisos del cuerpo adquiere aún más importancia.	0.480	0.748
According to Gerhard Friedrich, “runaway slaves who were caught used to be branded in their foreheads.	Según Gerhard Friedrich, “a los esclavos fugitivos que detenían los marcaban en la frente con hierro candente.	0.487	0.814
And vital to obedience is godly fear, yes, fear of displeasing God.	Y para obedecer es vital que tengamos temor piadoso, sí, temor de desagradar a Dios.	0.377	0.704
These large areas of “conflicting regional interests” were often the consequence of military conquest, since kings were invariably military leaders.	Dichas zonas extensas con ‘intereses regionales contrapuestos’ solían obedecer a conquistas, pues los reyes siempre eran caudillos militares.	0.423	0.712
If the criminal willingly conforms to punitive orders and exhibits changes in his attitude and behavior, a judge or president may choose to pardon him by lessening his sentence or totally forgiving his sentence.	En algunos países, los jueces y algunos funcionarios de alto rango tienen potestad para conmutar la pena a un delincuente —o incluso indultarlo— si este acepta su castigo y demuestra buena conducta.	0.087	0.654
In order to make ends meet, some mothers turn to prostitution and sell illegal drugs or encourage their daughters to do so.	Para subsistir, algunas se prostituyen y trafican con drogas, o empujan a sus hijas a tales actividades.	0.247	0.574
Our power of reason also tends to shy away from things that seem hopelessly vague and undefinable.	La razón también tiende a descartar todo lo que parece totalmente vago e indefinible.	0.407	0.690
When the first tremor subsided, they rushed outside and saw each other. Together they ran to higher ground.	Cuando pasó el terremoto, salieron corriendo de sus casas, se encontraron en la calle y huyeron a un lugar alto.	0.303	0.632
Indeed, kind parents strive to discern what discipline works best for each of their children.	Sin duda, los padres bondadosos tratan de descubrir cuáles son las medidas que funcionan mejor con cada hijo.	0.487	0.674
Others favor retaining the three-line form, making the middle one slightly longer.	Otros prefieren conservar la forma de tres versos, pero alargan un poco el segundo.	0.337	0.710
She used to bring substantial profit to her masters by fortune-telling.	Era una esclava, y ganaba mucho dinero para sus dueños, adivinando.	0.250	0.566
Later on, though, I began to feel that maybe I had been a little extreme.	Pero luego me puse a pensar si no me habría ido al otro extremo.	0.390	0.706

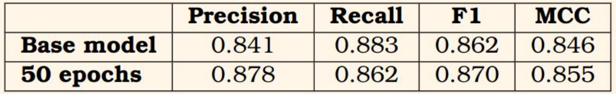
Improving Bicleaner AI

Increasing epochs

Base Model







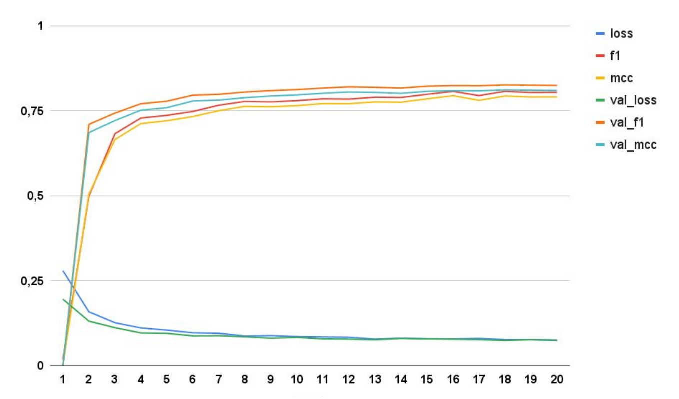
	Precision	**Recall**	**F1**	**MCC**
Base model	0.841	0.883	0.862	0.846
50 epochs	0.878	0.862	0.870	0.855

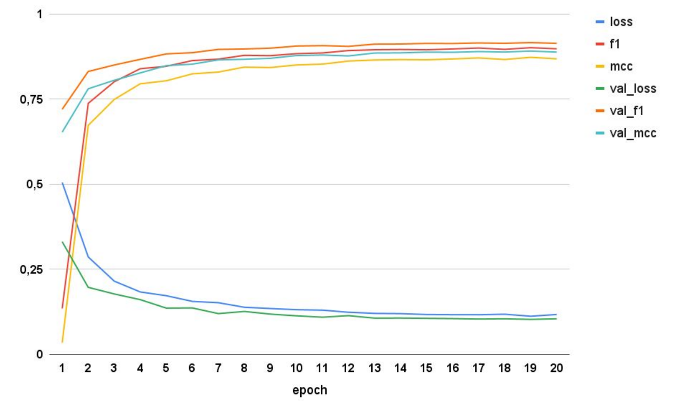
Improving Bicleaner AI

Noisy samples parameters

Noise generation:

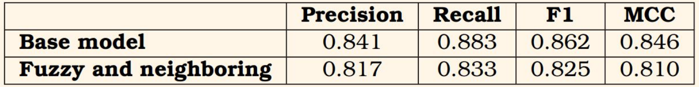
- Sentence modification
- Misalignment sentences
  - `pos_ratio` & `rand_ratio`:
- Frequence based
  - `freq_ratio` & `min_freq_words`
- Word Omision
  - `womit_ratio` & `womit_ratio`
- Fuzzy Matching
  - `fuzzy_ratio`
- Neighboring Sentence Misalignment
  - `neighbour_mix`



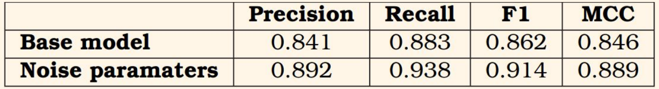


8421s ~ 2h 20min 21s

14765s ~ 4h 6min 5s



	Precision	Recall	F1	MCC
Base model	0.841	0.883	0.862	0.846
Fuzzy and neighboring	0.817	0.833	0.825	0.810

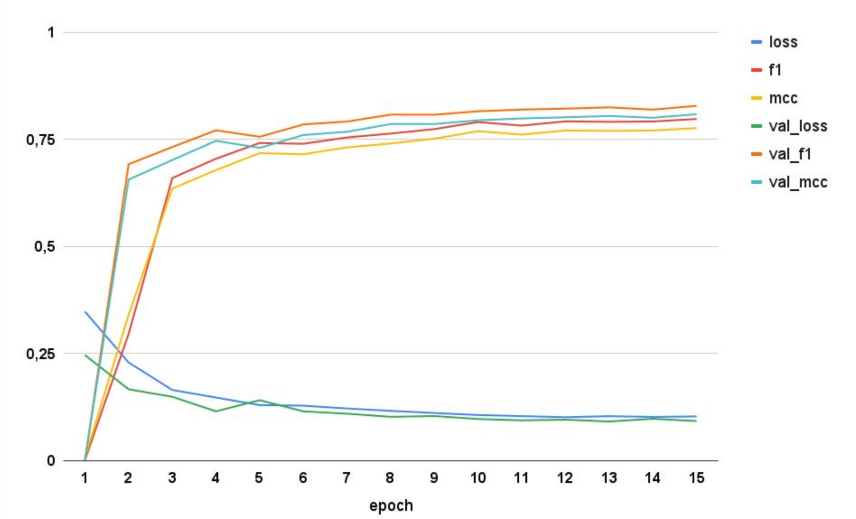


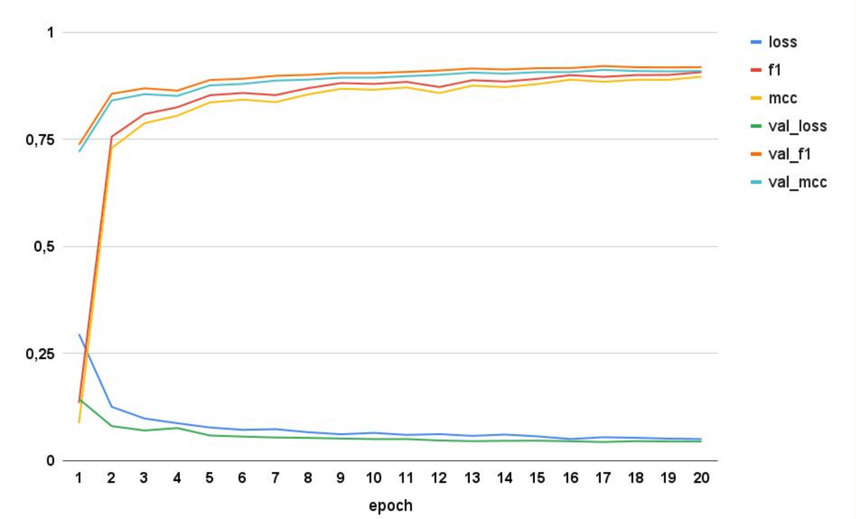
	Precision	Recall	F1	MCC
Base model	0.841	0.883	0.862	0.846
Noise paramaters	0.892	0.938	0.914	0.889

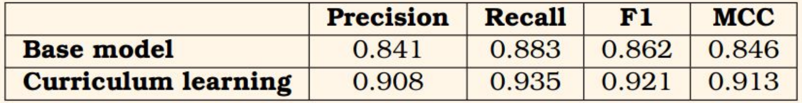
Improving Bicleaner AI

Curriculum learning

Base Model







	Precision	Recall	F1	MCC
Base model	0.841	0.883	0.862	0.846
Curriculum learning	0.908	0.935	0.921	0.913

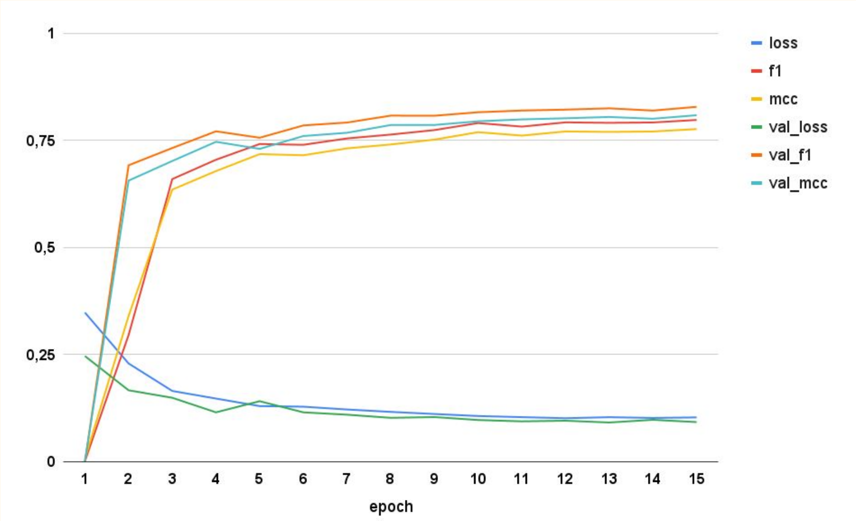
Improving Bicleaner AI

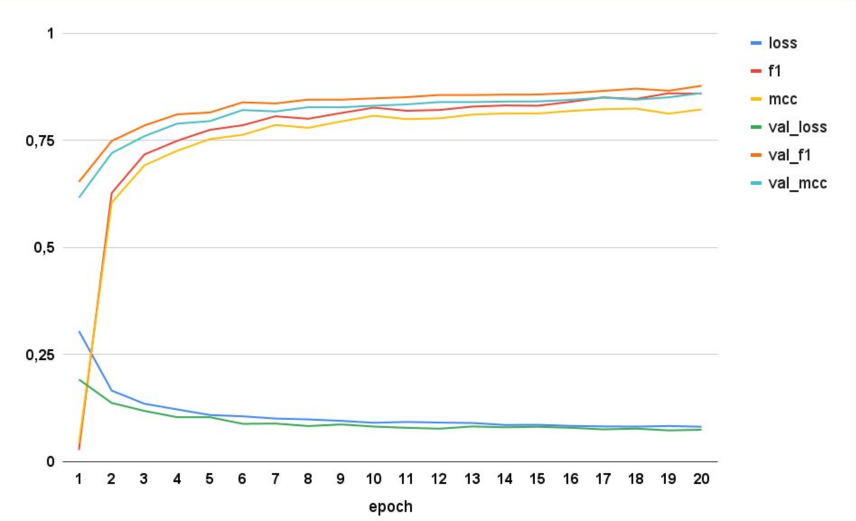
Hyperparameter tuning

Units in the hidden layer

	Precision	Recall	F1	MCC
Base model (2048)	0.841	0.883	0.862	0.846
N_hidden 4096	0.885	0.881	0.888	0.869

Base Model



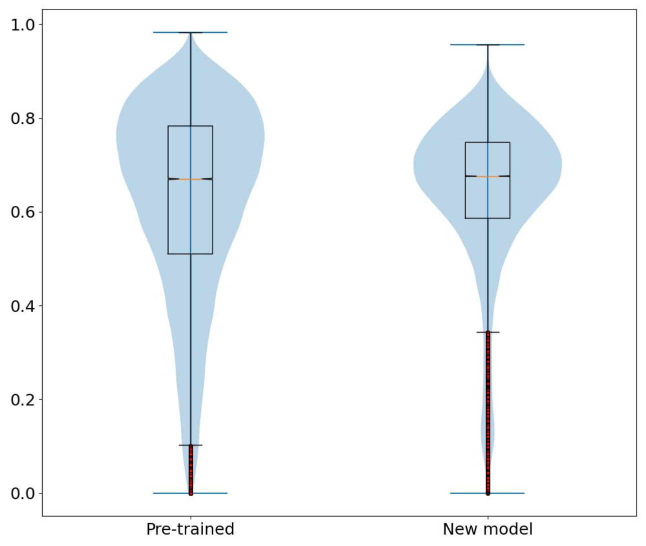


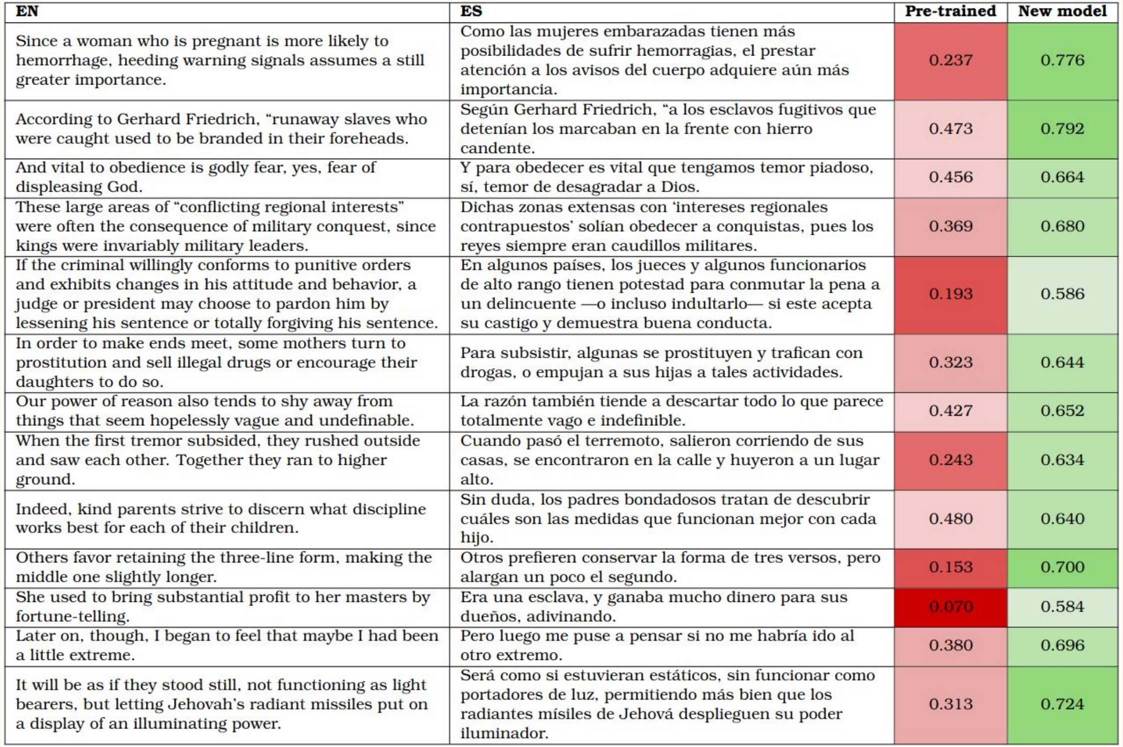
04

Results

Final models

Bicleaner classic



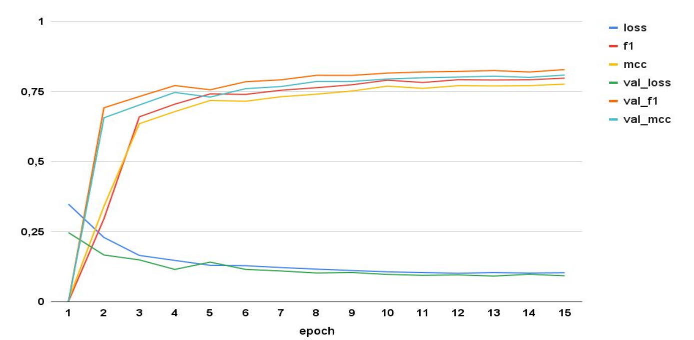


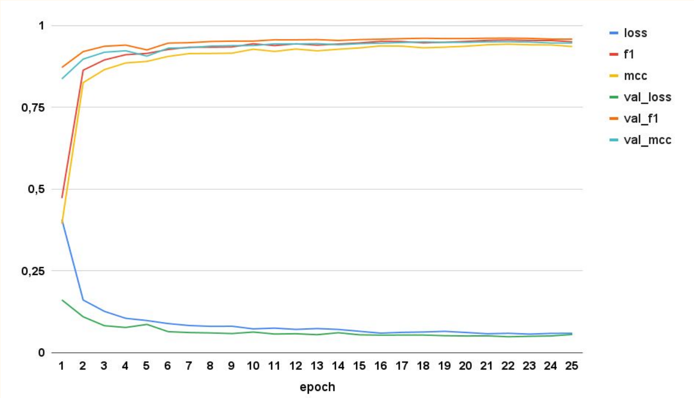
EN	ES	Pre-trained	New model
Since a woman who is pregnant is more likely to hemorrhage, heeding warning signals assumes a still greater importance.	Como las mujeres embarazadas tienen más posibilidades de sufrir hemorragias, el prestar atención a los avisos del cuerpo adquiere aún más importancia.	0.237	0.776
According to Gerhard Friedrich, "runaway slaves who were caught used to be branded in their foreheads.	Según Gerhard Friedrich, "a los esclavos fugitivos que detenían los marcaban en la frente con hierro candente.	0.473	0.792
And vital to obedience is godly fear, yes, fear of displeasing God.	Y para obedecer es vital que tengamos temor piadoso, sí, temor de desagradar a Dios.	0.456	0.664
These large areas of "conflicting regional interests" were often the consequence of military conquest, since kings were invariably military leaders.	Dichas zonas extensas con 'intereses regionales contrapuestos' solían obedecer a conquistas, pues los reyes siempre eran caudillos militares.	0.369	0.680
If the criminal willingly conforms to punitive orders and exhibits changes in his attitude and behavior, a judge or president may choose to pardon him by lessening his sentence or totally forgiving his sentence.	En algunos países, los jueces y algunos funcionarios de alto rango tienen potestad para conmutar la pena a un delincuente —o incluso indultarlo— si este acepta su castigo y demuestra buena conducta.	0.193	0.586
In order to make ends meet, some mothers turn to prostitution and sell illegal drugs or encourage their daughters to do so.	Para subsistir, algunas se prostituyen y trafican con drogas, o empujan a sus hijas a tales actividades.	0.323	0.644
Our power of reason also tends to shy away from things that seem hopelessly vague and undefinable.	La razón también tiende a descartar todo lo que parece totalmente vago e indefinible.	0.427	0.652
When the first tremor subsided, they rushed outside and saw each other. Together they ran to higher ground.	Cuando pasó el terremoto, salieron corriendo de sus casas, se encontraron en la calle y huyeron a un lugar alto.	0.243	0.634
Indeed, kind parents strive to discern what discipline works best for each of their children.	Sin duda, los padres bondadosos tratan de descubrir cuáles son las medidas que funcionan mejor con cada hijo.	0.480	0.640
Others favor retaining the three-line form, making the middle one slightly longer.	Otros prefieren conservar la forma de tres versos, pero alargan un poco el segundo.	0.153	0.700
She used to bring substantial profit to her masters by fortune-telling.	Era una esclava, y ganaba mucho dinero para sus dueños, adivinando.	0.070	0.584
Later on, though, I began to feel that maybe I had been a little extreme.	Pero luego me puse a pensar si no me habría ido al otro extremo.	0.380	0.696
It will be as if they stood still, not functioning as light bearers, but letting Jehovah's radiant missiles put on a display of an illuminating power.	Será como si estuvieran estáticos, sin funcionar como portadores de luz, permitiendo más bien que los radiantes mísiles de Jehová desplieguen su poder iluminador.	0.313	0.724

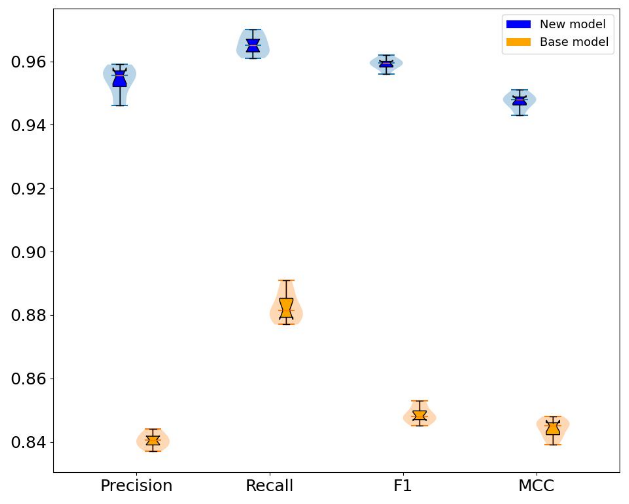
Final models

Bicleaner AI full

Base Model







	Precision	Recall	F1	MCC
Base model	0.841±0.0007	0.883±0.0014	0.862±0.0008	0.846±0.0009
New model	0.953±0.0015	0.965±0.0010	0.959±0.0005	0.947±0.0007

05

Conclusions

Conclusions

Bicleaner classic

- Wider corpus

Bicleaner AI full

- Increasing epochs
- Curriculum learning
- Noise parameters
- Increasing units in the hidden layers

¡Muchas gracias!