Universitat Politècnica de Catalunya

FACULTAT D'INFORMÀTICA DE BARCELONA

PRACTICA 3 - EXTRACCIÓ D'ENTITATS ANOMENADES

Processament del Llenguatge Humà

Grau en Intel·ligència Artificial

Andreu López Pérez
Pablo Chacón Martín

17 de maig de 2025

ÍNDEX

1. Introducció

1.1 Objectiu

En aquesta pràctica s'ha desenvolupat un sistema de reconeixement d'entitats anomenades (Named Entity Recognition, NER) mitjançant models de Conditional Random Fields (CRF), utilitzant la llibreria nltk.tag.CRFTagger de Python. L'objectiu principal és entendre com diferents estratègies d'enginyeria de features i codificacions d'etiquetes influeixen en el rendiment d'un model seqüencial per a la detecció d'entitats en text.

El projecte s'estructura en diverses fases, partint d'un model bàsic que només té en compte la paraula actual, fins a arribar a un model més complet que incorpora informació morfològica, contextual, llistes de paraules i embeddings semàntics. Al llarg del desenvolupament, s'han entrenat models per a dues llengües (espanyol i neerlandès) emprant el corpus CoNLL-2002, i també s'han realitzat proves sobre textos reals.

Finalment, s'analitza l'evolució del model, es discuteixen els resultats obtinguts tant quantitativament com qualitativament, i es proposen possibles línies de millora, incloent-hi l'extensió opcional amb el corpus CADEC.

1.2 Recursos i preprocessat

Per a aquesta pràctica s'ha utilitzat el corpus CoNLL-2002, accessible amb NLTK tan per els conjunts de entrenament, desenvolupament i test del espanyol com el neerlandès. Els conjunts de dades s'han transformat a una llista de tuplas (paraula, etiqueta), descartant el POS inicialment per a la versió base.

2. Model Base amb CRFTagger

2.1 Implementació sense features

La primera versió del sistema es troba al notebook NER_sense_features.ipynb. Aquesta implementació busca establir una base de rendiment utilitzant únicament la funcionalitat per defecte del CRFTagger de NLTK, sense incorporar cap mena de característiques addicionals (features). L'objectiu és obtenir una línia base de comparació per futures millores del model.

Es fa ús del corpus CoNLL-2002, que ja està preparat per a la tasca de reconeixement d'entitats (NER) tant en espanyol com en neerlandès. Aquest corpus proporciona frases en format BIO (Inside, Outside, Beginning), i el codi processa cada frase deixant només tuples de la forma (paraula, etiqueta), descartant altres anotacions com les etiquetes morfosintàctiques.

Amb aquesta configuració, el model CRF només té accés a la seqüència plana de paraules i la seva etiqueta associada. No pot explotar informació com la capitalització, la presència de signes de puntuació, les posicions relatives dins la frase ni cap mena d'element morfològic o contextual. Això limita molt la capacitat del model per detectar patrons útils, però permet avaluar fins a quin punt les etiquetes es poden predir únicament des de la seqüència observada.

2.2 Resultats i anàlisis d'errors

Per avaluar el rendiment del sistema s'ha emprat la mètrica F1-Score, ja que proporciona un equilibri entre precisió (precision) i cobertura (recall), dues dimensions essencials en tasques de reconeixement d'entitats. A diferència de l'exactitud (accuracy), que pot resultar enganyosa en certs casos, el F1-Score captura millor la capacitat real del sistema per identificar entitats rellevants (tal com es va comentar a classe).

En el cas de l'espanyol, el model ha obtingut un F1-Score ponderat del 70.38%, mentre que per al neerlandès el resultat ha estat del 63.11%. Aquest rendiment inferior en neerlandès pot atribuir-se en part a un menor volum d'entitats en el conjunt d'entrenament, així com a una distribució menys homogènia de les etiquetes.

A la Taula 1 es mostren els valors de precisió, recall i F1 per cada etiqueta (espanyol), que permeten veure amb més detall el comportament del sistema:

Reporte por clase:

	precision	recall	f1-score	support
B-LOC	0.74	0.68	0.71	1084
B-MISC	0.74	0.45	0.56	339
B-ORG	0.78	0.81	0.79	1400
B-PER	0.78	0.79	0.78	735
I-LOC	0.60	0.54	0.57	325
I-MISC	0.53	0.47	0.50	557
I-ORG	0.79	0.79	0.79	1104
I-PER	0.83	0.94	0.88	634
O	0.00	0.00	0.00	277
accuracy			0.70	6455
macro avg	0.64	0.61	0.62	6455
weighted avg	0.71	0.70	0.70	6455

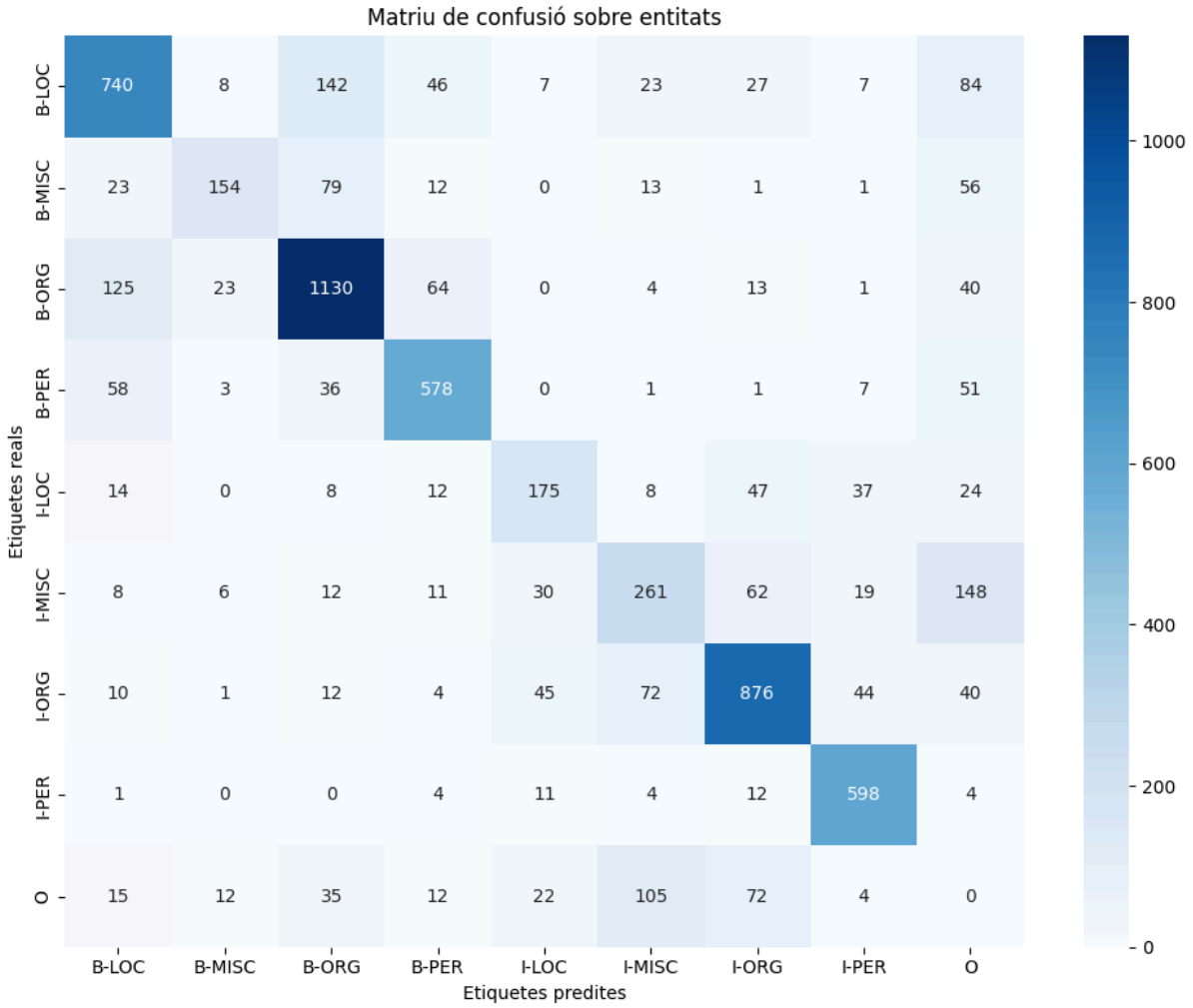
En paral·lel, la Taula 2 mostra el nombre de prediccions correctes i errònies per cada etiqueta, així com el percentatge d'encerts:

	Etiqueta	Correctes	Errors	% d'encerts
0	B-LOC	740	254	74.45%
1	B-MISC	154	53	74.40%
2	B-ORG	1130	324	77.72%
3	B-PER	578	165	77.79%
4	I-LOC	175	115	60.34%
5	I-MISC	261	230	53.16%
6	I-ORG	876	235	78.85%
7	I-PER	598	120	83.29%
8	O	0	447	0.00%

Finalment, la Taula 3 mostra la distribució de freqüències de les etiquetes en el conjunt d'entrenament:

	Etiqueta	Freqüència
0	O	231920
1	B-ORG	7390
2	I-ORG	4992
3	B-LOC	4913
4	B-PER	4321
5	I-PER	3903
6	I-MISC	3212
7	B-MISC	2173
8	I-LOC	1891

L'anàlisi d'aquestes dades revela que els millors resultats es donen en entitats com I-PER i B-ORG, mentre que les etiquetes I-MISC i I-LOC presenten els pitjors percentatges d'encerts. En particular, l'etiqueta O (que indica que una paraula no forma part de cap entitat) sovint es confon amb etiquetes d'entitat, fet que es reflecteix clarament a la matriu de confusió:



En el cas del model entrenat per al neerlandès, els resultats són similars però amb una caiguda notable del F1-Score. A continuació, es mostren les dades equivalents:

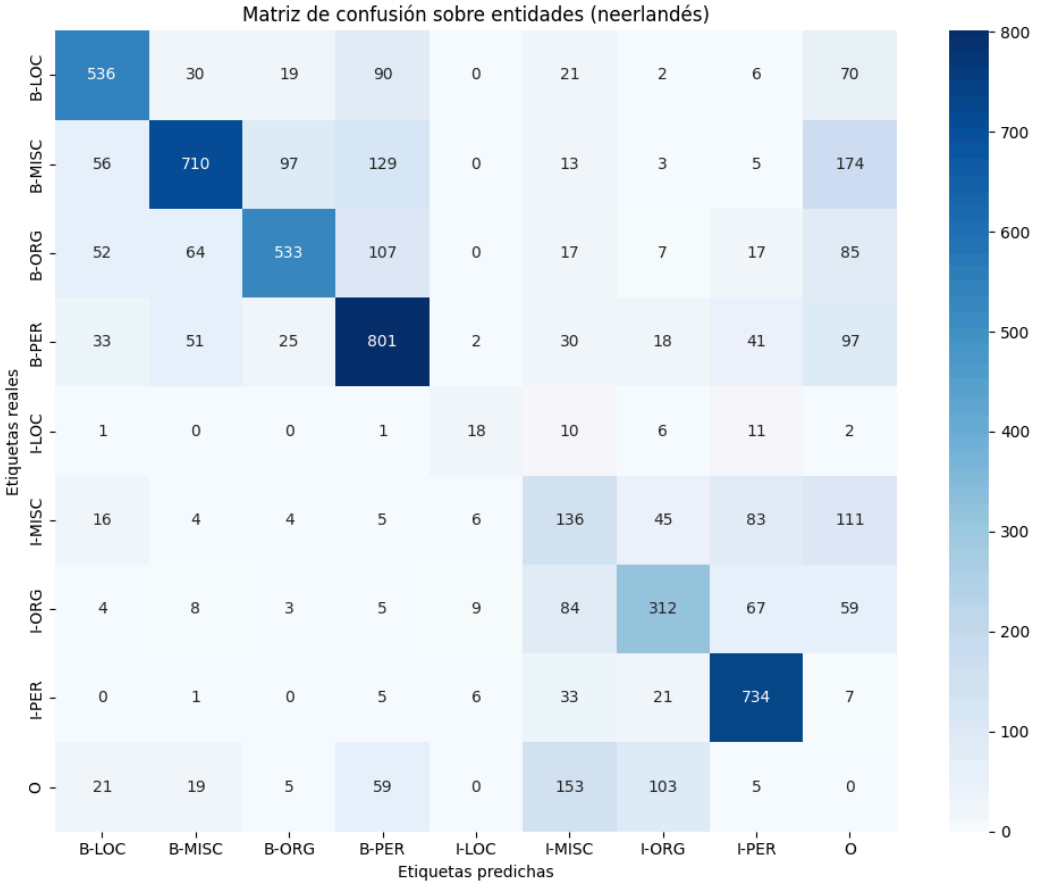
	precision	recall	f1-score	support
B-LOC	0.75	0.69	0.72	774
B-MISC	0.80	0.60	0.68	1187
B-ORG	0.78	0.60	0.68	882
B-PER	0.67	0.73	0.70	1098
I-LOC	0.44	0.37	0.40	49
I-MISC	0.27	0.33	0.30	410
I-ORG	0.60	0.57	0.58	551
I-PER	0.76	0.91	0.83	807
O	0.00	0.00	0.00	365
accuracy			0.62	6123
macro avg	0.56	0.53	0.54	6123
weighted avg	0.66	0.62	0.63	6123

Rendiment per etiqueta:

	Etiqueta	Correctes	Errors	% d'encerts
0	B-LOC	536	183	74.55%
1	B-MISC	710	177	80.05%
2	B-ORG	533	153	77.70%
3	B-PER	801	401	66.64%
4	I-LOC	18	23	43.90%
5	I-MISC	136	361	27.36%
6	I-ORG	312	205	60.35%
7	I-PER	734	235	75.75%
8	O	0	605	0.00%

Freqüència d'etiquetes en entrenament (neerlandés):

	Etiqueta	Freqüència
0	O	183346
1	B-PER	4716
2	B-MISC	3338
3	B-LOC	3208
4	I-PER	2883
5	B-ORG	2082
6	I-MISC	1405
7	I-ORG	1199
8	I-LOC	467



Tot i que hi ha una correlació parcial entre la freqüència d'una etiqueta en el conjunt d'entrenament i el seu rendiment, aquest factor no ho explica tot. Per exemple, I-PER en espanyol obté un 83% d'encerts amb una freqüència moderada, mentre que I-MISC, amb una presència similar, no arriba al 55%. Això suggereix que no només la quantitat d'exemples sinó també la claredat dels patrons contextuals i morfològics associats a cada entitat són determinants.

Aquestes observacions reforcen la hipòtesi que el model es beneficiaria significativament de la incorporació de features contextuals, com la capitalització, la posició dins la frase o l'estructura de la paraula. Aquest serà l'objectiu de la següent etapa de desenvolupament.

3. Model amb extracció de features

En la següent etapa, l'objectiu és dotar el model de més coneixement sobre la morfologia i el context de les paraules. Aquesta estratègia es desenvolupa en els scripts: NER_amb_features_1/2/3/4.ipynb, que representen les diferents generacions de millora.

3.1 Primer model

3.1.1 Implementació

En aquest primer intent de millora, es defineix una funció de features explícita que, per a cada paraula, afegeix informació rellevant per a la tasca de NER. Les característiques que s'extrauen inclouen:

- Si la paraula comença amb majúscula.
- Si conté números.
- Si està formada només per signes de puntuació.
- Diversos sufixos i prefixos (fins a longitud 3).
- La paraula anterior i la següent (context local).

Aquestes features es basen en la intuïció que, per exemple, moltes entitats comencen amb majúscula, que els sufixos com "-ción" o "-dad" poden indicar organitzacions, i que el context de la paraula pot ajudar a desambiguar.

3.1.2 Resultats i anàlisis d'errors

Aquest primer model amb features mostra una millora substancial respecte al model base. El F1-Score (weighted) puja fins al **76.41%**, millorant gairebé 6 punts percentuals respecte al model sense cap mena de característica contextual o morfològica.

La taula següent mostra el detall de precisió, recall i F1 per cada classe. Les millores són especialment notables en entitats com **B-LOC**, **B-ORG** i **I-PER**, mentre que es mantenen dificultats en entitats més ambigües o poc representatives com **I-MISC** o **O**.

Reporte por clase:

	precision	recall	f1-score	support
B-LOC	0.82	0.77	0.79	1084
B-MISC	0.71	0.50	0.59	339
B-ORG	0.81	0.85	0.83	1400
B-PER	0.85	0.87	0.86	735
I-LOC	0.71	0.61	0.65	325
I-MISC	0.62	0.57	0.59	557
I-ORG	0.82	0.81	0.81	1104
I-PER	0.90	0.94	0.92	634
O	0.00	0.00	0.00	203
accuracy			0.76	6381
macro avg	0.69	0.66	0.67	6381
weighted avg	0.77	0.76	0.76	6381

A continuació es resumeix el rendiment per etiqueta, destacant el nombre de prediccions correctes i errònies, així com el percentatge d'encerts. Es pot veure com entitats com **I-PER** assoleixen gairebé un **90%** d'encert, mentre que altres com **I-MISC** encara mostren rendiments inferiors al 65%, probablement per la seva heterogeneïtat i baixa freqüència d'entrenament.

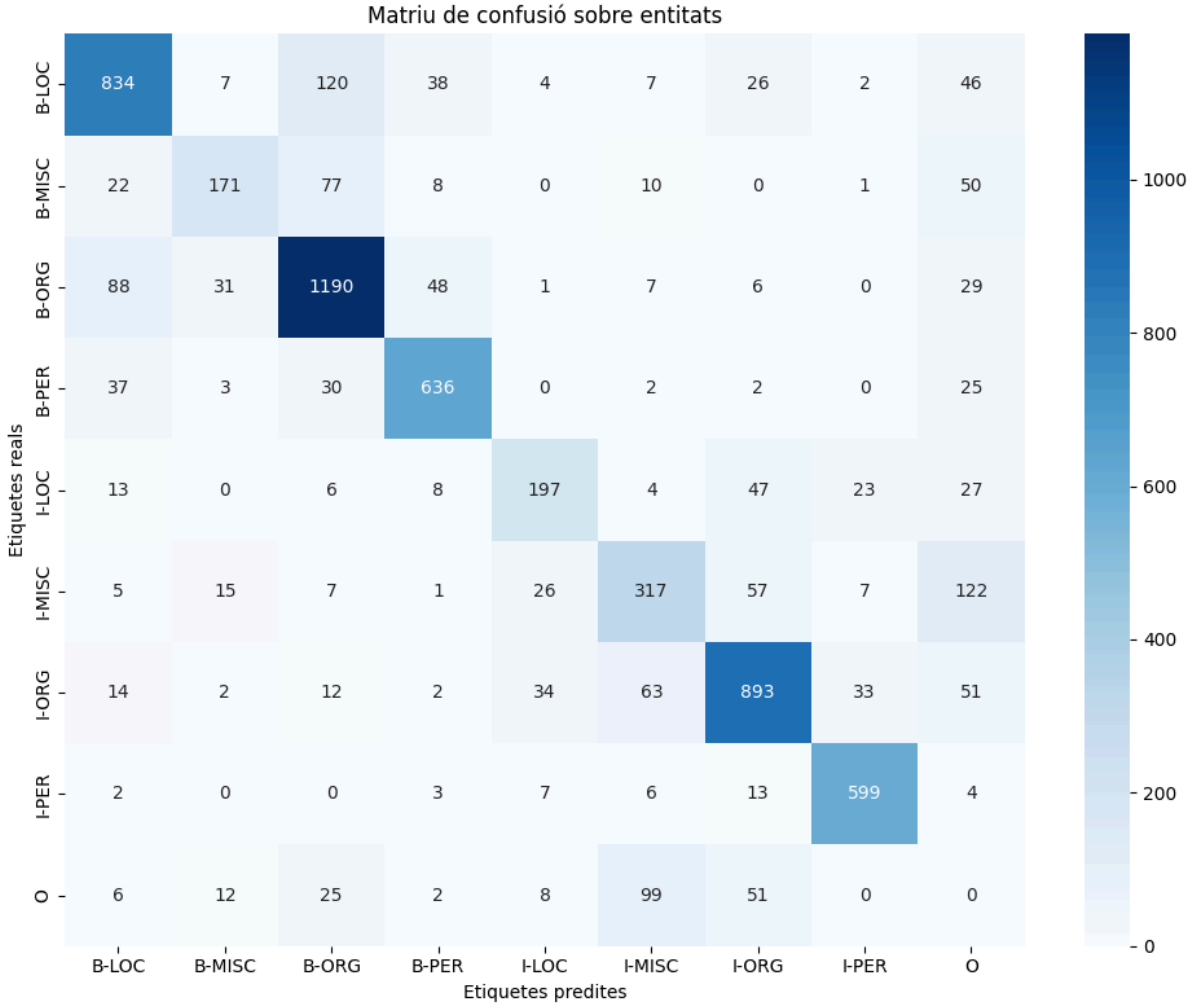
	Etiqueta	Correctes	Errors	% d'encerts
0	B-LOC	834	187	81.68%
1	B-MISC	171	70	70.95%
2	B-ORG	1190	277	81.12%
3	B-PER	636	110	85.25%
4	I-LOC	197	80	71.12%
5	I-MISC	317	198	61.55%
6	I-ORG	893	202	81.55%
7	I-PER	599	66	90.08%
8	O	0	354	0.00%

Aquestes millores reforcen la hipòtesi que el model es beneficia clarament de disposar de més informació contextual. Característiques com saber si una paraula comença amb majúscula o quines paraules l'envolten ajuden a detectar entitats amb més fiabilitat i a evitar falsos positius.

Cal destacar que poden semblar presents certes confusions notables amb l'etiqueta "O", que acumula una gran quantitat d'errors i cap encert, però que no s'ha de tenir en compte ja que realment no ho són. No és que no encertem en cap cas l'etiquetatge de "O", sinó que

les que són TRUE NEGATIVE no ens interessen per l'anàlisi d'errors. De fet la gran majoria de 'O' realment hem comprobat que sí que les encerta.

A continuació es mostre la matriu de confusió referent a aquest primer model:



Aquest primer experiment valida la direcció adoptada: l'ús de features pot augmentar significativament la capacitat predictiva del model. En els següents models, s'explorarà com optimitzar encara més aquest conjunt de característiques.

3.2 Segon model

3.1.1 Implementació

Després de validar que les features bàsiques proporcionen una millora substancial, en aquesta fase es decideix ser més ambiciós i es prova d'afegir-ne encara més:

- Prefixos i sufixos fins a longitud 4, per captar patrons més llargs.

- Diferenciació més fina de la capitalització (per exemple, si es troba a inici de frase).
- Ús de stopwords per a filtrar paraules que habitualment no són entitats.
- S'experimenta amb l'ús d'etiquetes morfosintàctiques (POS-tags), tot i que es detecta que això fa més lent l'entrenament i les prediccions.
- Es deixa preparada la possibilitat d'usar llistes de paraules (gazetteers), tot i que no s'inclouen en la versió final per simplicitat.

3.1.2 Resultats i anàlisis d'errors

Aquest segon model representa un pas més en la millora del sistema, amb l'addició de features més elaborades. Tot i no incorporar encara fonts externes com *gazetteers*, les noves característiques —com prefixos i sufixos més llargs, distincions més fines en la capitalització, i ús de *stopwords*— permeten captar millor patrons morfològics i posicionals.

El F1-Score (weighted) millora lleugerament fins al **76.83%**, donant una mica a entendre que el model es podria haver estabilitzat en una zona de rendiment òptim amb les dades disponibles i el conjunt actual de característiques.

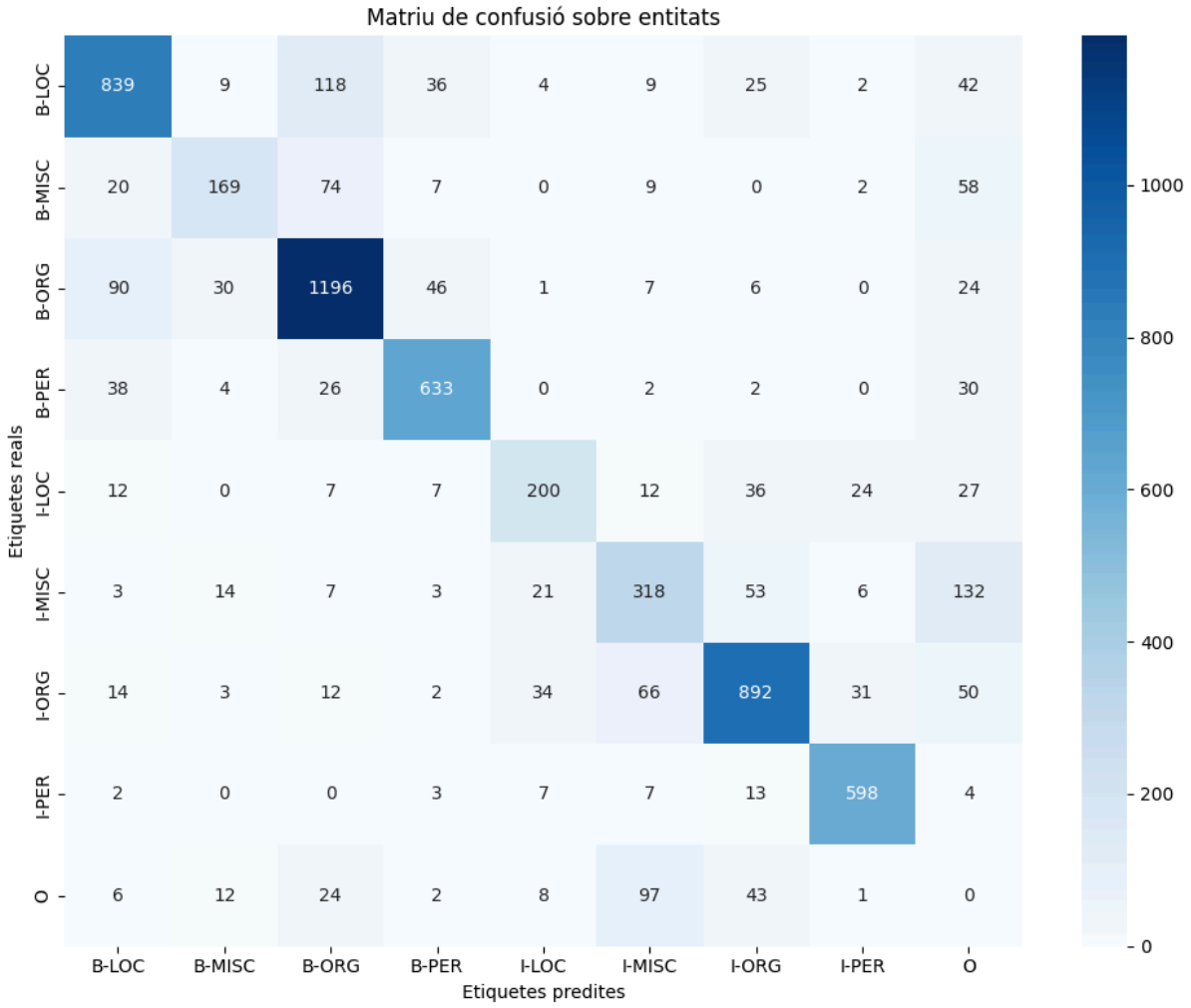
Reporte por clase:

	precision	recall	f1-score	support
B-LOC	0.82	0.77	0.80	1084
B-MISC	0.70	0.50	0.58	339
B-ORG	0.82	0.85	0.84	1400
B-PER	0.86	0.86	0.86	735
I-LOC	0.73	0.62	0.67	325
I-MISC	0.60	0.57	0.59	557
I-ORG	0.83	0.81	0.82	1104
I-PER	0.90	0.94	0.92	634
O	0.00	0.00	0.00	193
accuracy			0.76	6371
macro avg	0.70	0.66	0.67	6371
weighted avg	0.78	0.76	0.77	6371

Les millores respecte al primer model són modestes, però consistents, sobretot en etiquetes com **I-ORG** i **I-LOC**, que incrementen lleugerament el seu F1. També es manté un alt rendiment en **I-PER**, amb un F1 superior al 90%.

	Etiqueta	Correctes	Errors	% d'encerts
0	B-LOC	839	185	81.93%
1	B-MISC	169	72	70.12%
2	B-ORG	1196	268	81.69%
3	B-PER	633	106	85.66%
4	I-LOC	200	75	72.73%
5	I-MISC	318	209	60.34%
6	I-ORG	892	178	83.36%
7	I-PER	598	66	90.06%
8	O	0	367	0.00%

I també es mostra a continuació la matriu de confusió per a aquest segon cas:



Tot i que els guanys són moderats, aquest segon model confirma que la qualitat i especificitat de les features tenen un impacte positiu en el rendiment. També obre la porta a considerar altres millores, com l'ús efectiu de *gazetteers*, embeddings o tècniques d'ensemblatge per continuar refinant la detecció d'entitats.

3.3 Tercer model (afegir embeddings)

3.1.1 Implementació

Un cop s'ha assolit el límit pràctic de millora amb features morfològiques i lèxiques, es decideix fer un salt qualitatiu (tot i saber que els embeddings formen part de la pràctica 4 i no d'aquesta): incorporar coneixement semàntic a través d'embeddings de paraula, ja que es creu que d'aquesta manera es pot assolir una detecció més favorable a partir de les relacions semàntiques. Aquesta estratègia es desenvolupa al notebook NER_amb_features_3.ipynb. També s'ha elaborat en aquest cas el notebook corresponent al neerlandès per veure, arribats ja a un bon punt, com afecten les millores proposades a l'altra llengua a estudiar. Tots els resultats pertinents es poden veure en el notebook tot i no ser comentats de forma explícita en el report.

Els embeddings, com els vectors de Word2Vec, permeten representar cada paraula com un vector numèric que captura la seva similitud semàntica amb altres paraules. Per exemple, "Madrid" i "Barcelona" tindran vectors propers, de manera que el model pot generalitzar millor a entitats desconegudes però semànticament properes a les del corpus d'entrenament. En el nostre cas hem fet servir **models preentrenats** descarregats desde https://vectors.nlpl.eu/repository/ amb els corresponents **ID's 39 (neerlandès) i 68 (espanyol)**.

La funció de features s'amplia per incloure, per a cada paraula, els primers components del vector embedding, a més de totes les característiques morfològiques i de context ja utilitzades.

3.1.2 Resultats i anàlisis d'errors

Amb la incorporació d'embeddings, el model fa un salt qualitatiu en la seva capacitat de generalització. Els vectors semàntics aporten informació que no depèn només de la forma o del context local, sinó de les relacions latents entre paraules. Això fa que el model pugui reconèixer entitats menys freqüents o noves que, tot i no haver estat vistes en entrenament, són semànticament properes a d'altres conegudes.

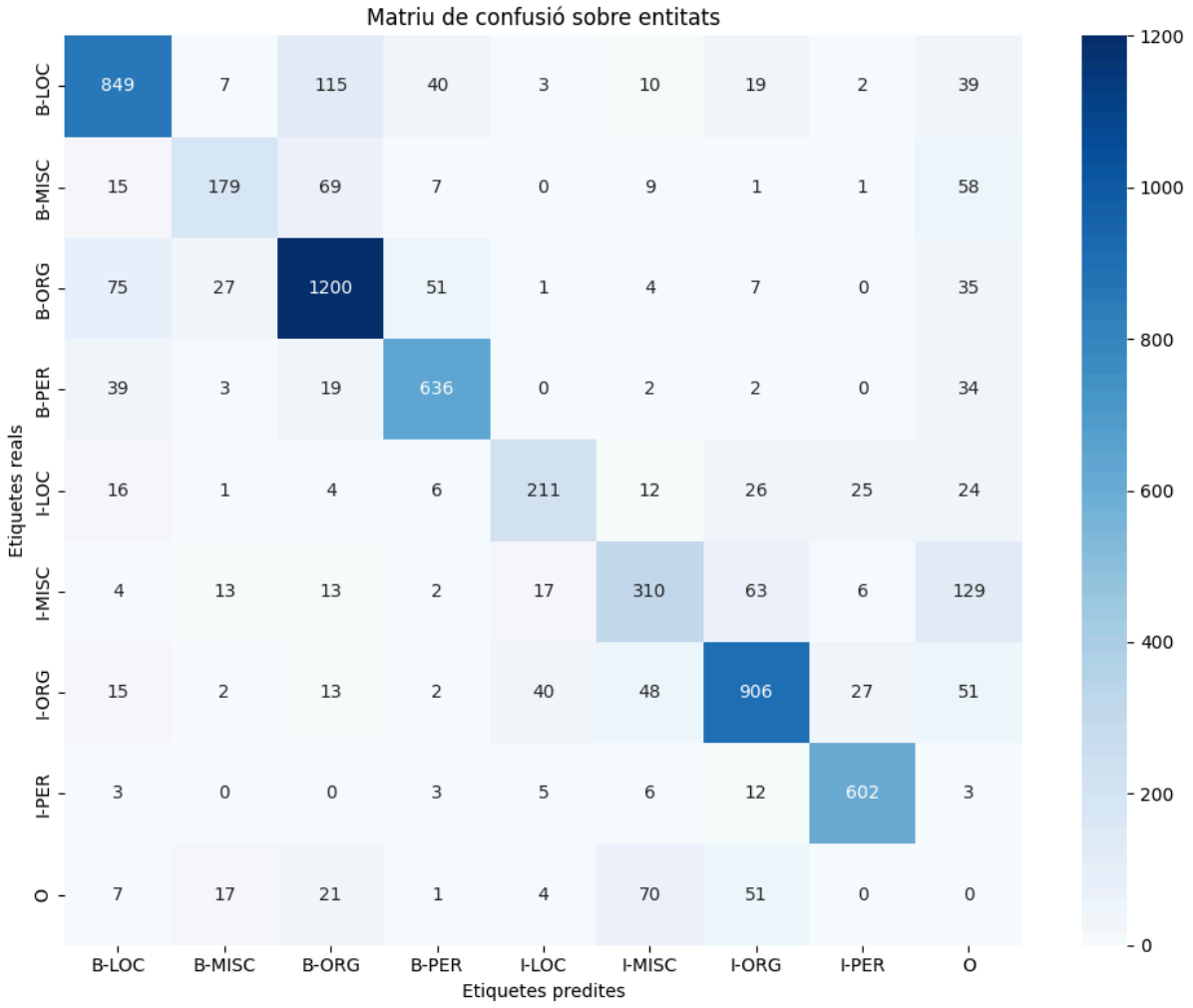
El F1 Score (weighted) millora fins al **77.99%**, mostrant un increment substancial respecte als models anteriors. Aquesta millora es distribueix de forma consistent en gairebé totes les etiquetes, amb especial relleu en les etiquetes compostes com **I-LOC** i **I-MISC**, que tendeixen a ser més difícils de capturar.

Reporte por clase:

	precision	recall	f1-score	support
B-LOC	0.83	0.78	0.81	1084
B-MISC	0.72	0.53	0.61	339
B-ORG	0.83	0.86	0.84	1400
B-PER	0.85	0.87	0.86	735
I-LOC	0.75	0.65	0.70	325
I-MISC	0.66	0.56	0.60	557
I-ORG	0.83	0.82	0.83	1104
I-PER	0.91	0.95	0.93	634
O	0.00	0.00	0.00	171
accuracy			0.77	6349
macro avg	0.71	0.67	0.69	6349
weighted avg	0.79	0.77	0.78	6349

	Etiqueta	Correctes	Errors	% d'encerts
0	B-LOC	849	174	82.99%
1	B-MISC	179	70	71.89%
2	B-ORG	1200	254	82.53%
3	B-PER	636	112	85.03%
4	I-LOC	211	70	75.09%
5	I-MISC	310	161	65.82%
6	I-ORG	906	181	83.35%
7	I-PER	602	61	90.80%
8	O	0	373	0.00%

Es confirma que l'etiqueta **I-PER** continua sent una de les més fàcilment reconegudes, amb un F1 del 93% i un percentatge d'encerts superior al 90%. En canvi, l'etiqueta **I-MISC** continua sent la que presenta menor percentatge d'encert, classificant la majoria d'entitats mal etiquetades, com a "**O**". Tal com s'aprecia un cop més en la següent matriu de confusió igual que en les anteriors.



De totes maneres, aquest resultat valida la utilitat dels embeddings per a la tasca de NER, especialment en corpus amb certa riquesa lèxica i diversitat d'entitats. També assenta les bases per a possibles futures integracions amb embeddings contextuals (com els de BERT), o per a la combinació amb *gazetteers* per afinar encara més les prediccions, tot i que no s'aplicarà en aquesta pràctica, ja que considerem que seria enarça'n massa del plantejament de la pràctica fet.

3.4 Quart model (clustering amb embeddings)

3.1.1 Implementació

Ara bé, sí que s'ha volgut implementar una última idea, tenint en compte que la incorporació d'embeddings al nostre model va incrementar significativament el temps d'entrenament i predicció, probablement degut a l'alta dimensionalitat afegida. Per abordar aquest problema, vam explorar estratègies per reduir la dimensionalitat sense perdre capacitat predictiva. La solució adoptada va ser aplicar clustering amb K-means als vectors d'embedding, utilitzant només l'ID del cluster com a feature.

Al model NER_amb_features_4, vam reemplaçar els 10 valors inicials dels embeddings per aquest enfocament més eficient. Aquesta optimització:

- Redueix dràsticament la dimensionalitat (de 10 a 1 feature)
- Agrupa paraules amb significats similars en els mateixos clusters
- Manté la capacitat de generalització semàntica
- Millora el rendiment computacional

Aquest són els resultats d'alguns clústers després de l'entrenament:

```
Palabras en el clúster 232:
Antonio
Francisco
Fernando
Jesús
Sánchez
Pedro
López
Rodríguez
Josep
Alberto
Denis
Vázquez
Rafael
```

```
Palabras en el clúster 95:
Madrid
Barcelona
San
Oviedo
Badajoz
Santander
Santiago
Málaga
Sevilla
Cantabria
Bilbao
Sebastián
Janeiro
```

Per als embeddings, vam utilitzar FastText entrenat exclusivament amb les nostres dades d'entrenament. Això permet gestionar paraules no vistes anteriorment inferint-ne el significat a partir dels seus fragments (n-grames).

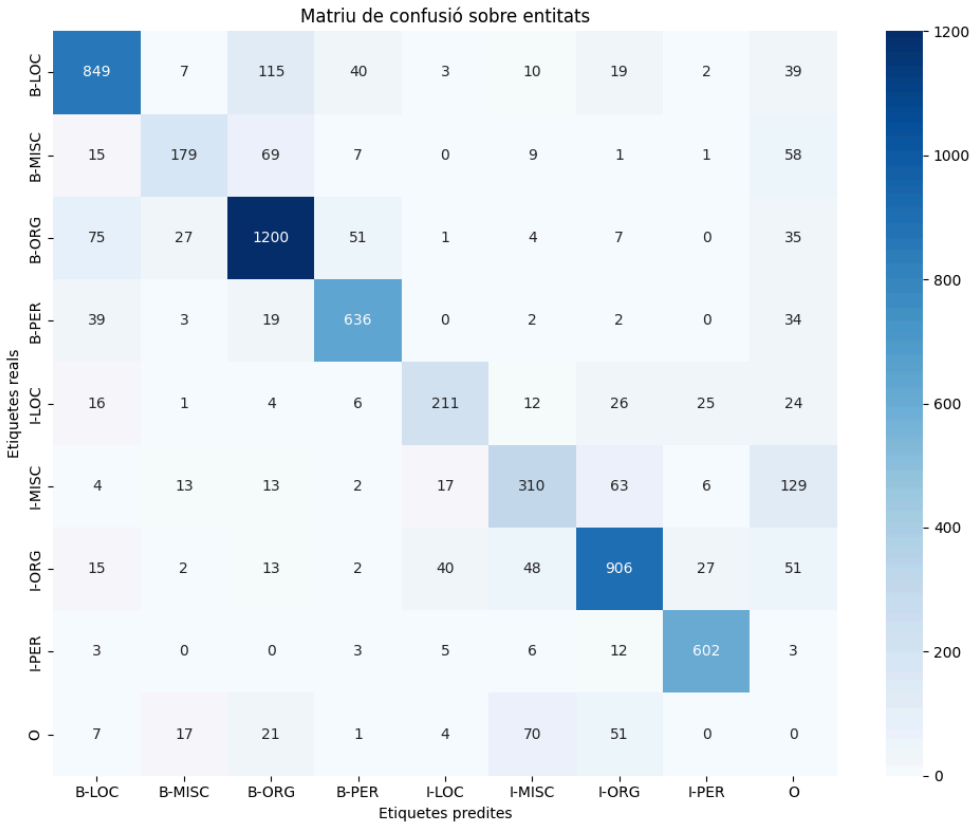
3.1.2 Resultats i anàlisis d'errors

Amb aquests canvis el model no aconsegueix tan bons resultats com en el model 3:

```
Classification Report:
              precision    recall  f1-score   support

       B-LOC     0.7998    0.7408    0.7692      1084
      B-MISC     0.6603    0.5103    0.5757       339
       B-ORG     0.8106    0.8436    0.8267      1400
       B-PER     0.8382    0.8531    0.8456       735
       I-LOC     0.7003    0.6185    0.6569       325
      I-MISC     0.5938    0.5853    0.5895       557
       I-ORG     0.8248    0.7844    0.8041      1104
       I-PER     0.8770    0.9338    0.9045       634
           O     0.0000    0.0000    0.0000       212

    accuracy                         0.7463      6390
   macro avg     0.6783    0.6522    0.6636      6390
weighted avg     0.7616    0.7463    0.7529      6390
```



La reducció en el rendiment respecte al model anterior (del 77.99% al 75.29% en F1-score ponderat) probablement s'atribueix a la pèrdua d'informació semàntica en aplicar clustering als embeddings. Tot i que el K-means redueix la dimensionalitat (de 10 valors a 1 cluster ID), aquesta pot ser una simplificació massa excessiva que no és capaç de diferenciar tant bé entre paraules amb significats similars però no idèntics, afectant especialment categories complexes com MISC.

També perd la precisió dels vectors continus que capturaven millor les relacions semàntiques subtils i pot generar clusters poc definits per a entitats amb poca representació en el training. Les categories més genèriques (com PER) es mantenen estables per la seva consistència semàntica, mentre que les més específiques o variables (LOC, MISC) pateixen per la falta de detall en les representacions. Això suggereix que, en aquest cas, el balanç entre eficiència i precisió s'ha inclinat massa cap a la simplificació.

Els nostres experiments van mostrar que incrementant progressivament el nombre de clústers (k) s'obtenien millores mínimes en el rendiment del model fins a assolir un llindar als aproximadament 300 clústers, punt en què les millores s'estabilitzaven. Aquest comportament suggereix que existeix un límit en el benefici que pot obtenir-se mitjançant aquesta estratègia de clustering.

Teòricament, podríem entrenar els embeddings FastText amb un corpus més extens i diversificat, equilibrant millor les dades per a cada tipus d'entitat. Segurament així

obtindríem millors resultats, no obstant això, donat que el model 3 (amb embeddings complets) ja ofereix millors resultats, hem decidit no aprofundir més en aquesta línia d'investigació, centrant els nostres esforços en experimentar amb diferents codificacions.

4. Model Optimitzat

4.1.1 Implementació

En l'última fase del projecte, el focus es desplaça cap a una optimització sistemàtica dels hiperparàmetres i la comparació de diferents esquemes de codificació d'etiquetes, tot aprofitant el conjunt de features més ric i efectiu identificat en les fases anteriors.

El notebook *NER_grid_search_amb_codificacions.ipynb* exemplifica aquest procés, aportant un enfocament exhaustiu i metòdic per a la selecció de la millor configuració possible per al reconeixement d'entitats. El procediment es pot dividir en dues parts principals: Recerca de la millor configuració de features (utilitzant BIO) i comparació de les diferents codificacions d'etiquetes.

Primerament, es realitza una cerca exhaustiva (grid search) sobre els hiperparàmetres que controlen la inclusió de diferents tipus de features:

- Embeddings (vectors semàntics de Word2Vec)
- Context (si es considera la paraula anterior/següent)
- Puntuació (presència de signes de puntuació)

S'utilitza la codificació BIO (Begin, Inside, Outside) com a referència per determinar quina combinació de features proporciona els millors resultats de F1-score.
El grid search avalua totes les combinacions possibles d'activació/desactivació d'aquestes característiques, entrenant un model per a cadascuna i avaluant-lo sobre el conjunt de test. El procés és automàtic i objectiu, assegurant que la selecció dels hiperparàmetres sigui òptima i no arbitrària.

El millor conjunt de features identificat per BIO (al notebook, per exemple: embeddings i context activats, puntuació desactivada) es pren com a "configuració òptima" per a la següent fase.

Amb la millor configuració de features ja establerta, s'aplica aquest mateix conjunt de característiques a diversos esquemes de codificació d'entitats:

- BIO (Begin-Inside-Outside)

- IO (Inside-Outside)
- BIOW (BIO + Word, per a paraules soles)
- BIOES (Begin-Inside-Outside-End-Single)

Per cadascuna d'aquestes codificacions, es torna a entrenar el model i s'avalua de manera rigorosa sobre el mateix conjunt de test. Això permet aïllar l'efecte de la codificació de les etiquetes respecte als resultats, ja que la resta de condicions (features, embeddings, etc.) es mantenen constants.

4.1.2 Resultats i anàlisis d'errors

La recerca sistemàtica de la millor combinació de features revela que desactivar la puntuació, mantenint els embeddings i el context, permet millorar lleugerament el rendiment. Això suggereix que els signes de puntuació poden introduir soroll o reduir la capacitat de generalització del model en aquesta tasca concreta.

Un cop fixada aquesta configuració, es compara el rendiment del model sota diferents esquemes de codificació d'etiquetes, mantenint constants totes les altres variables. L'objectiu és entendre com afecta la manera de representar les entitats a la precisió del reconeixement.

Tabla comparativa:

	Precision	Recall	F1-Score
BIO	0.792606	0.771267	0.780346
IO	0.792606	0.771267	0.780346
BIOW	0.774790	0.758057	0.764702
BIOES	0.767499	0.748384	0.756110

Els resultats mostren que tant l'esquema BIO com IO ofereixen els millors F1-score (0.780), mentre que BIOW i BIOES presenten un rendiment lleugerament inferior. És destacable que la codificació IO, tot i ser menys informativa (no distingeix entre inicis i interiors d'entitats), arriba al mateix nivell de rendiment, cosa que suggereix que el model és capaç de compensar aquesta mancança gràcies a les altres features.

D'altra banda, esquemes més elaborats com BIOES, que teòricament haurien de proporcionar més granularitat, no semblen aportar una millora pràctica. Això podria deure's al fet que l'increment en complexitat no es tradueix en una major capacitat de generalització, i potser fins i tot dificulta l'aprenentatge.

Aplicant el mateix al neerlandès obtenim els següents resultats:

```
Tabla comparativa:
       Precision    Recall  F1-Score
BIO     0.751763  0.718433  0.732280
IO      0.751763  0.718433  0.732280
BIOW    0.743899  0.709156  0.721436
BIOES   0.743652  0.707176  0.717797
```

On es pot veure que també es dona el cas on les codificacions BIO i IO són les més altes totes dues amb els mateixos valors i tenint en compte que els resultats per al neerlandès respecte al script *NER_amb features_3_ned.ipynb* eren d'un F1-Score final de **0.7285**. Podem destacar que en aquest cas gràcies a l'optimitzador, també millora mínimament assolint un valor final de **0.7322**.

Aquestes observacions permeten extreure que el guany principal en aquesta etapa final prové més de l'optimització dels hiperparàmetres que no pas del canvi d'esquema de codificació, i que codificacions més simples com BIO continuen sent una opció sòlida i eficient per a models basats en features.

5. Comparació i anàlisi final

5.1 Comparació de resultats

Els diferents models desenvolupats al llarg del projecte mostren una evolució clara pel que fa a la seva capacitat de reconèixer entitats. Des del primer enfocament basat únicament en característiques morfològiques simples fins a l'últim, que integra embeddings semàntics i optimització sistemàtica, s'observa una millora sostinguda del rendiment.

Model	F1 Score (weighted)
Model amb features bàsics	**76.41%**
Model amb features ampliats	**76.83%**
Model amb embeddings	**77.99%**
Model optimitzat (BIO)	**78.03%**

Les millores no són sempre espectaculars en magnitud absoluta, però sí consistents, cosa que demostra l'impacte acumulatiu de les bones decisions de disseny. L'última versió, que aplica grid search i compara esquemes de codificació, obté el millor rendiment tot mantenint la simplicitat relativa del model (CRF) i aprofitant de manera eficient els recursos disponibles.

5.2 Similituds en Errors d'Identificació

Una revisió transversal dels errors entre models revela algunes tendències comunes:

- Entitats compostes o amb noms poc freqüents segueixen sent un repte. Tant si són llocs menys coneguts com organitzacions amb noms llargs o amb ambigüitats lèxiques, aquestes entitats tenen un comportament erràtic entre prediccions.

- MISC és, de lluny, la classe més problemàtica, amb baixos valors de recall en tots els models. Aquesta etiqueta, tendeix a contenir entitats molt diverses i difícils de definir.

- Els errors també augmenten quan hi ha canvis subtils en el context, especialment quan les entitats apareixen en posicions gramaticals menys habituals (per exemple, dins d'incisos o en frases molt curtes).

Aquestes similituds indiquen que hi ha límits estructurals en el model, probablement deguts a l'arquitectura seqüencial i a la manca de coneixement extern específic (com gazetteers o informació sintàctica profunda).

5.3 Conclusió i possibles millores

El projecte posa de manifest que els models de Conditional Random Fields (CRF), malgrat la seva simplicitat comparativa, poden oferir resultats molt competitius si se'ls dota d'un conjunt de features acuradament seleccionades i optimitzades. L'evolució del model mostra clarament que no només cal afegir més informació, sinó saber quina informació és rellevant i com combinar-la de manera eficient.

L'ús d'embeddings semàntics representa un punt d'inflexió: permet al model generalitzar millor i capturar relacions entre paraules que no comparteixen forma però sí significat. Això obre la porta a l'ús de models més sofisticats basats en embeddings contextuals, com BERT o RoBERTa, que poden millorar encara més els resultats sense dependre tant de la selecció manual de features.

En particular, l'últim script desenvolupat destaca com a la culminació d'aquesta progressió:

- L'optimització garanteix l'eficiència i evita la sobreingesta de soroll.

- L'avaluació de diferents codificacions d'etiquetes demostra que esquemes simples com BIO o IO són sorprenentment robustos per a aquest tipus de dades.

- Es mostra que la clau per seguir millorant no és necessàriament la complexitat estructural, sinó una estratègia metòdica i guiada per dades.

Un cop realitzat el projecte, tot i ja haver estat comentat alguna vegada al llarg d'aquest, algunes possibles propostes de millora que hem pensat i podrian arribar a aplicar-se de cara al futur basant-nos en el punt final al que s'ha arribat podrien ser:

- Incorporar embeddings contextuals com BERT o XLM-R, especialment en un entorn multilingüe.

- Augmentar la quantitat i varietat del corpus d'entrenament, incloent-hi més entitats de domini específic.

- Utilitzar gazetteers i recursos lingüístics externs per millorar la cobertura d'entitats rares.

- Explorar features sintàctiques (com dependències gramaticals) que puguin ajudar a la desambiguació d'entitats en contextos complexos.

6. EXTRA - Corpus de CADEC

El CADEC Corpus (Consumer Adverse Drug Event Corpus) és un corpus anotat de comentaris, que conté experiències personals relacionades amb l'ús de medicaments i els seus efectes adversos.

La seva forma és la d'un fitxer amb format tipus BIO (Beginning–Inside–Outside), on cada paraula d'un comentari apareix en una línia separada, seguida d'etiquetes que indiquen si forma part d'una entitat biomèdica específica. Cada comentari està identificat per un codi de fàrmac (com CATAFLAM.2 o LIPITOR.108) i s'anota quines paraules són l'inici (B-) o continuació (I-) d'una entitat, juntament amb un codi SNOMED CT que representa conceptes mèdics estandarditzats.

Un exemple seria:

LIPITOR.108

Muddled	B-10012805	O	O	O	O
thinking	I-10012805	O	O	O	O
,	O	O	O	O	O
loss	B-10041909	O	O	O	O
of	I-10041909	O	O	O	O
strength	I-10028350	O	O	O	O
and	O	O	O	O	O
stamina	I-10041909	O	O	O	O
.	O	O	O	O	O

Per poder entrenar un model de reconeixement d'entitats amb el CRFTagger de la llibreria NLTK, cal que les dades estiguin en un format específic: una llista de frases, on cada frase és una llista de tuples del tipus (paraula, etiqueta_NER). Per aquest motiu, implementem un parser que llegeix el corpus original (en format .conll) i transforma les dades en aquest format estructurat. El parser identifica on comença una nova frase, extreu les etiquetes NER corresponents a cada paraula (seguint el format BIO), i organitza tota la informació de manera que sigui directament usable pel CRFTagger.

Així quedaria el mateix exemple vist abans:

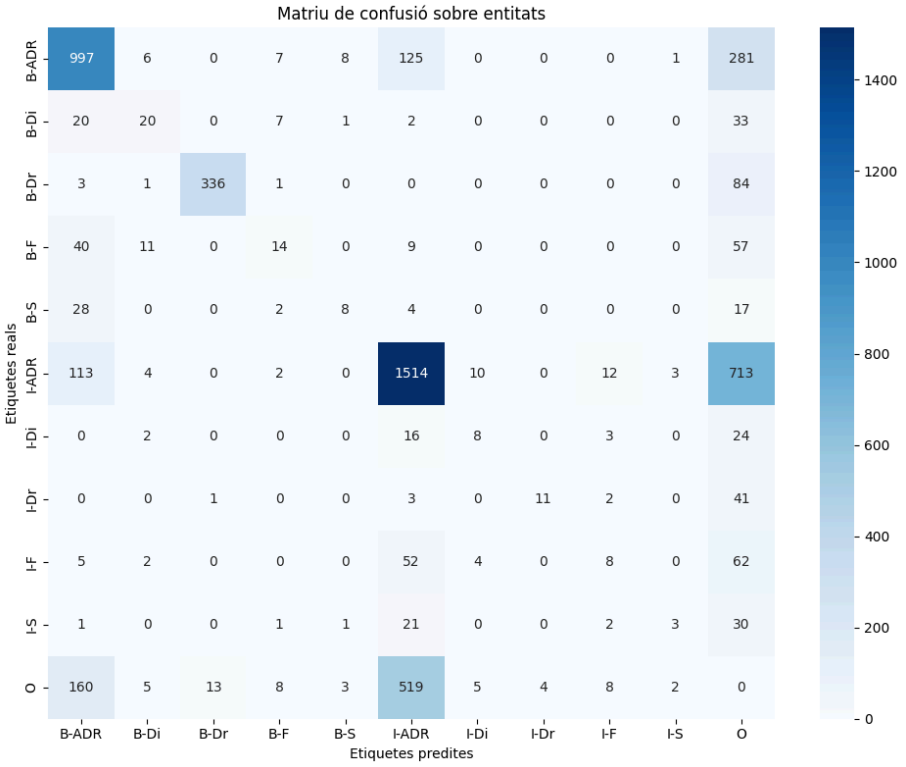
[('Muddled', 'B-ADR'), ('thinking', 'I-ADR'), (',', 'O'), ('loss', 'B-ADR'), ('of', 'I-ADR'), ('strength', 'I-ADR'), ('and', 'O'), ('stamina', 'I-ADR'), ('.', 'O')]

Fet això vam realitzar el entrenament amb els features del model 3 (NER_amb_features_3) on vam utilitzar el model de words embedings preentrenats amb anglés de https://vectors.nlpl.eu/repository/ (**ID 40**).

Aquests són els resultats del primer intent:

Classification Report:

	precision	recall	f1-score	support
B-ADR	0.7293	0.6996	0.7142	1425
B-Di	0.3922	0.2410	0.2985	83
B-Dr	0.9600	0.7906	0.8671	425
B-F	0.3333	0.1069	0.1618	131
B-S	0.3810	0.1356	0.2000	59
I-ADR	0.6684	0.6385	0.6531	2371
I-Di	0.2963	0.1509	0.2000	53
I-Dr	0.7333	0.1897	0.3014	58
I-F	0.2286	0.0602	0.0952	133
I-S	0.3333	0.0508	0.0882	59
O	0.0000	0.0000	0.0000	727
accuracy			0.5284	5524
macro avg	0.4596	0.2785	0.3254	5524
weighted avg	0.5864	0.5284	0.5501	5524



Les entitats més freqüents, com I-ADR (7582 exemples) i B-ADR (4506), obtenen f1-scores raonables de 0.65 i 0.71 respectivament, cosa que indica que el model aprèn millor quan hi ha prou dades. En canvi, entitats com B-F, I-F, B-S, I-S o I-Di,

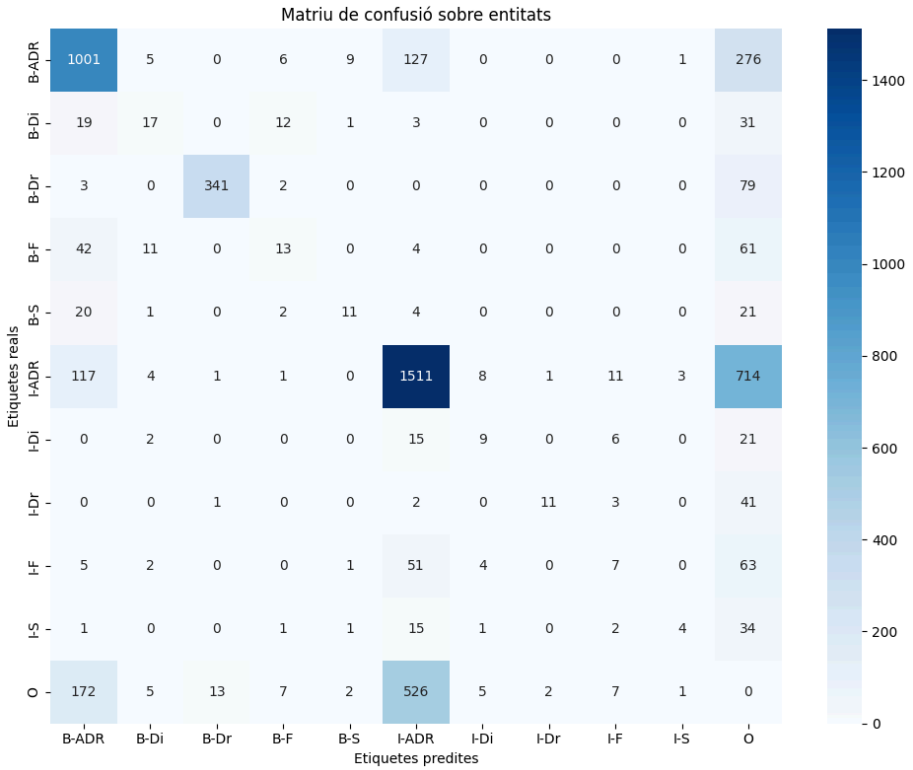
amb menys de 300 exemples cadascuna, tenen f1-scores molt baixos (sovint per sota de 0.2), fet que reflecteix la dificultat del model per generalitzar amb entitats poc representades. També es veu que les etiquetes interiors (I-) acostumen a obtenir pitjors resultats que les inicials (B-), fet que pot indicar problemes per capturar correctament la continuïtat de les entitats. En conjunt, l'accuracy global del 52.8% i la mitjana ponderada de f1-score del 0.55 reflecteixen un rendiment acceptable, amb molt marge de millora especialment en la detecció d'entitats menys freqüents.

Donat el fort desbalanceig de dades en l'entrenament —amb entitats com B-F, I-S o I-Di molt poc representades— el model té dificultats per reconèixer correctament certes etiquetes. Per compensar aquesta manca de dades, hem incorporat una sèrie de features lingüístiques i semàntiques dissenyades per ajudar el model a identificar patrons propis del llenguatge mèdic. Entre aquestes característiques s'inclouen aspectes formals com sufixos mèdics comuns (-itis, -oma, etc.); més tokens de context: dues paraules anteriors i posteriors. També s'identifiquen tokens que contenen caràcters especials (-, /, ()).

Amb aquests nous features aquest és el resultat:

Classification Report:

	precision	recall	f1-score	support
B-ADR	0.7254	0.7025	0.7137	1425
B-Di	0.3617	0.2048	0.2615	83
B-Dr	0.9579	0.8024	0.8732	425
B-F	0.2955	0.0992	0.1486	131
B-S	0.4400	0.1864	0.2619	59
I-ADR	0.6692	0.6373	0.6528	2371
I-Di	0.3333	0.1698	0.2250	53
I-Dr	0.7857	0.1897	0.3056	58
I-F	0.1944	0.0526	0.0828	133
I-S	0.4444	0.0678	0.1176	59
O	0.0000	0.0000	0.0000	740
accuracy			0.5283	5537
macro avg	0.4734	0.2830	0.3312	5537
weighted avg	0.5847	0.5283	0.5491	5537



Tot i haver afegit un conjunt de features específiques per captar millor tecnicismes mèdics (com sufixos mèdics, presència de símbols, capitalització, embeddings, etc.), el rendiment global del model no ha millorat. L'accuracy es manté al mateix nivell i l'F1-score ponderat baixa lleugerament. Això indica que, encara que aquestes features poden ser útils teòricament, no són suficients per compensar el fort desbalanceig del conjunt d'entrenament, on les entitats com B-ADR, I-ADR o B-Dr tenen una presència molt més gran que la resta.

La major part de les entitats minoritàries (com I-F, I-S, B-S, etc.) continuen tenint valors de recall i F1 molt baixos, i fins i tot en alguns casos empitjoren lleugerament. Per exemple:

- B-Di baixa de 0.2985 a 0.2615 d'F1.
- I-Dr millora molt poc: de 0.3014 a 0.3056.

El problema principal continua sent el desbalanceig de les etiquetes. Sense més exemples per a les entitats menys freqüents o sense tècniques de reequilibri (com oversampling, undersampling o pèrdues ponderades), l'addició de features no resulta suficient per millorar el rendiment del model.