Universitat de les Illes Balears

Escola Politècnica Superior

Grau en Enginyeria Informàtica

Aprenentatge Automàtic - 21781

Pràctica 1 - SVM

Professor:

Biel Moyà Alcover

Alumnes:

Jonathan Salisbury Vega

Julià Wallis Medina

Índex

Introducció al problema

Aquesta pràctica consisteix en desenvolupar un model de classificació que permeti classificar l'idioma d'una paraula en base a las característiques de la propia paraula. Es demana que el model estigui basat en SVM, però es deixa a lliure elecció el tipus de kernel i parametrització del model.

Les dades que es dona consisteixen en 1000 paraules en català i anglès.

Solucions considerades

Procés a seguir

1. **Modificacions del dataset:** El primer que s'ha és augmentar el dataset afegint 8 nous idiomes (Alemany, Francès, Espanyol, Italià, Polonès, Portuguès, Rus i Suec). Afegint aquests idiomes hi ha 5 llengües romàniques, 2 germàniques, 2 eslaves i 1 nòrdica.

   També s'ha hagut de netejar les dades, ja que l'eina de traducció que s'ha emprat a vegades traduïa les paraules erroneament (*Ang:* Twenty -> *Esp:* 20)
2. **Creació de característiques:** S'ha enfocat la creació de les característiques amb dos punts.

   S'han creat tant característiques "generals" com a característiques específiques per a cada idioma. S'aprofundirà més dins les característiques creades a la secció corresponent
3. **Creació de models:** Una vegada creades les característiques s'han separat les dades en el conjunt d'entrenament, validació i test. Després s'han creat diferents models SVM amb distints kernels i paràmetres.
4. **Avaluació dels models:** Una vegada s'han entrenat els models, s'avaluaran. S'han usat distintes mètriques per poder avaluar els models *(Accuracy, precision, top_k_accuracy, ...)*
5. **Optimització dels models**: Gràcies a les mètriques s'han fet modificacions als models originals. A més s'ha emprat l'eina GridSearchCV.

Dades

Com ja s'ha explicat, el nostre dataset consisteix en 1000 paraules traduïdes a diversos idiomes.

Les 5 primeres entrades del dataset:

Ang	Cat	Esp	Ale	Fra	Pol	Por	Rus	Ita	Sue
as	com	como	wie	comme	jak	como	в качестве	come	som
his	seva	su	seine	le sien	jego	seu	его	il suo	hans
that	que	que	das	ce	że	este	что	quello	den där
he	ell	él	er	il	on	ele	он	lui	han
was	era	estaba	war	a été	był	foi	был	er	var

Característiques

En total s'han creat 64 noves característiques, aquestes es poden dividir en quatre grups:

- Contament de caràcters
- Grups de caràcters
- Prefixes i Sufixes
- Llei de Zipf

Tenir més informació sobre les paraules serà útil a l'hora de predir l'idioma de les paraules.

S'ha de tenir en compte que gairebé totes les noves característiques son subconjunts de les característiques dels llenguatges. Per exemple, totes les paraules que contenen la lletra "ñ" pertanyen al espanyol, però no totes les paraules de l'espanyol contenen la lletra "ñ".
L'objectiu és el de crear prou característiques per a que conjuntament puguin cobrir la gran part de les característiques pròpies de cada idioma.

Contament de caràcters

Les característiques basades en el contament de caracters s'han dividit en dos grups:

- Característiques generals
- Característiques específiques a l'idioma

Característiques generals:
Les característiques generals extreuen informació sobre la pròpia paraula, com per exemple: el nombre de vocals, nombre d'accents, espais en blanc, ...

	Nom	**"Components"**
1	Longitud	Nombre de lletres
2	vocal	aeiou
3	accents	àèìòùáéíóú
4	accents_esq	àèìòù
5	accents_dre	áéíóú
6	dieresi	äëïöü
7	circumflex	âêîôû

8	Nombre de paraules	“espais en blanc”
9	Apostrofs	‘
10	Guions	-
11	Rares	kqwxyz

Característiques específiques a l’idioma
Aquestes característiques están enfocades específicament als 10 idiomes que tenim.

L’objectiu d’aquestes variables és que permetin identificar l’idioma de la paraula basant-se en els caràcters identificatius de cada idioma (els no comuns).

	Nom	**“Components”**
12	Enye	ñ
13	C trencada	ç
14	Eszett	ß
15	Polonés	ąćęłńóśźż
16	Português	ãõ
17	Rus	бвгджзийклмнпрст фцчшщъыьэюя
18	Suec	åäö
19	Espanyol	áéíóúüñ
20	Italià	àèéìíîòóùú
21	Francès	àâæçéèêëîïôœùûüÿ
22	Alemany	äöüß
23	Català	àèéíïòóúüç

Contament de grups de caràcters

S’han investigat quines son els grups de caràcters més comuns de cada idioma i després s’han contabilitzat el nombre de vegades que apareixen a les paraules.

També es contabiliza el nombre de parelles de vocals

	Nom	**“Components”**
24	grups_ang	sh, th, ch, ck, ...
25	grups_cat	ny, tx, sc, nc, ...
26	grups_esp	nd, nt, ch, rr, ...
27	grups_ale	tch, ck, ng, qu, ...
28	grups_por	tch, lh, nh, qu, ...
29	grups_pol	ch, dz, dł, di,
30	grups_ita	ch, gl, gn, sc, ...
31	grups_swe	ch, ck, cid, dt, ...
32	grups_fra	ch, che, eau, ent, ...
33	grups_rus	бл, вл, гл, дл, ...
34	diptongs	Parelles de les següents lletres: *aeiouàèìòùáéíóúäëïöüâêîôû*

Colcuns exemples dels resultats que s’obtindrien emprant els grups de l’anterior taula

Paraula	ang	cat	esp	ale	por	pol	ita	swe	fra	rus	Dipt.
menyscreure	0	2	0	0	0	0	1	0	0	0	1
главное	0	0	0	0	0	0	0	0	0	1	1

Prefixos i sufixos

S’han investigat quins son els prefixos i sufixos més comuns de cada idioma

Prefixos

	Nom	**“Components”**
35	pre_ang	anti, be, de, dis, ...
36	pre_esp	anti, auto, contra, des, ...

37	pre_cat	anti, ab, avant, arxi, ...
38	pre_ita	auto, dis, en, ex, ...
39	pre_fra	anti, auto, co, con, ...
40	pre_por	auto, co, contra, des, ...
41	pre_ale	be, ein, ent, er, ...
42	pre_swe	be, för, in, om, ...
43	pre_pol	przed, nad, na, pod, ...
44	pre_rus	анти, без, в, во, ...

Sufixos

	Nom	**“Components”**
45	suf_ang	able, al, ation, er, ...
46	suf_esp	ado, ador, aje, anza, ...
47	suf_cat	ana, aca, ada, al, ...
48	suf_ita	abile, are, ario, atore, ...
49	suf_fra	age, aille, ance, eau, ...
50	suf_por	al, ão, ar, ês, ...
51	suf_ale	bar, e, ei, er, ...
52	suf_swe	ande, are, bar, dom, ...
53	suf_pol	acja, ać, anie, eć, ...
54	suf_rus	больше, енький, ик, ичка, ...

Llei de Zipf

La llei de Zipf és una llei empírica que descriu que la freqüència de les paraules d’un llenguatge.
S’ha emprat la funció *zipf_frequency* de la llibreria *wordfreq*. Com es pot llegir a la documentació, la “freqüència Zipf” d’una paraula en un idioma és el logaritme en base 10 del nombre d’aparicions per cada mil milions de paraules en l’idioma indicat.

Per exemple: la paraula *seva* apareix en català 1 pic cada 500 paraules, així que el seu valor seria de 6.3

A la següent taula es poden veure les zipf_freqüències de cada paraula per als 10 idiomes:

Paraula	**ang**	**cat**	**esp**	**ale**	**por**	**pol**	**ita**	**swe**	**fra**	**rus**
be	6.79	4.21	4.75	4.27	4.59	4.47	4.72	5.04	4.51	4.09
casa	3.54	5.77	5.76	5.88	3.68	5.99	3.52	3.55	3.64	2.79
han	3.93	5.93	6.3	4.43	3.86	3.95	6.53	3.97	3.76	2.8

A la següent taula es pot veure la freqüència de cada paraula a cada idioma, és a dir, cada quan apareix cada paraula (*be* apareix 1 de cada 162 en angles, ...)

Paraula	**ang**	**cat**	**esp**	**ale**	**por**	**pol**	**ita**	**swe**	**fra**	**rus**
be	162	61660	17783	53703	25704	33884	19055	9120	30903	81283
casa	288403	1698	1738	1318	208930	1023	301995	281838	229087	1621810
han	117490	1175	501	37154	138038	112202	295	107152	173780	1584893

Models

En aquesta pràctica s'ha utilitzat només el model de aprenentatge supervisat Support Vector Machine (SVM) per a un problema de classificació multi-classe.
Aquest model es basa en la idea de trobar una separació òptima entre diferents classes en un espai de característiques.
L'SVM busca la línia de separació que tingui el màxim marge, és a dir, la distància més gran entre la línia de separació i els vectors de suport més propers de cada classe. Això és important perquè la distància entre les diferents classes és una mesura de la generalització del model, i una línia de separació amb un marge més gran serà menys sensible al soroll de les dades i, per tant, serà més robusta.

Com és ben conegut aquest model soporta diferents tipus de *kernel*:

- **Lineal:**
  És adequat per a separar dades linealment separables i és el més ràpid en termes de temps de càlcul.
- **Polynomial:**
  És adequat per a separar dades no linealment separables i es basa en la combinació de característiques amb potències diferents.

- **Radial (RBF):**
  És adequat per a separar dades no linealment separables i es basa en la similitud entre les dades amb relació a un punt central o "centre de masses".
- **Sigmoid:**
  És adequat per a separar dades no linealment separables i es basa en la funció sigmoide.

SVM no soporta directament classificació multiclasse, per això pot fer servir diversos enfocaments;

- **1 vs 1:** separa el problema en múltiples problemes de classificació binaria, un per cada parell de classes.
- **1 vs la-resta:** separa el problema en un conjunt de classificadors binaris, un per cada clase.

Posteriorment es realitzarà un conjunt de proves amb els *kernels* i metodes anteriors per intentar obtenir els millors resultats.

Per poder utilitzar les dades anteriors al model, primer s'han de separar en *target* (l'idioma) i *features* (les altres columnes).

A continuació s'han de dividir les dades en un conjunt d'entrenament, un de validació i un darrer de test. S'ha decidit dur a terme l'entrenament amb el 80% de les dades i dividir el 20% restant entre la validació i el test.
Dividir les dades en 3 conjunts es de gran importància. El conjunt d'entrenament, com diu el nom, servirà per poder entrenar el model. El conjunt de validació servirà per fer ajustaments als hiper-paràmetres dels models. Aquests ajustaments poden anar desde decidir el conjunt de característiques fins al kernel del model. Un pic es tengui el model final es podrà executar el model damunt les dades de test, aquestes dades seran completament noves per al model ja que no haurà sigut entrenat sobre elles, ni s'haurà "filtrat" informació a través dels hiper-paràmetres.

Una vegada separades les dades correctament s'ha decidit escalar-es. Això és degut a que el model SVM se basa en la distancia entre els diferents punts per poder crear l'hiperplà de separació.

Com s'ha explicat anteriorment el conjunt original no contenia característiques, i han sigut afegides dins aquest projecte. Visualitzant un poc les estadístiques de les caracteristiques se pot observar que la majoria es troba a un rang de dades paregut, així que les distàncies no haurien de ser molt diferents. Indiferentment s'ha decidit escalarles igualment degut a que s'ha considerat que no es perd informació i és més formalment correcte per a l'ús d'aquest tipus de model.

Finalment, també s'ha decidit aplicar una reducció de dimensionalitat a les dades, per tenir una menor quantitat de característiques i que sigui més fàcil treballar amb les dades. També existeix la possibilitat de que algunes de les característiques estiguin altament correlacionades i aplicant aquest mètode reduïm les redundancies.

Mètriques

Per avaluar el rendiment d'un model d'aprenentatge automàtic, és important utilitzar mètriques adequades que s'ajustin al problema en qüestió. En aquest cas, on s'està utilitzant un model de màquina de vectors de suport (SVM) per predir la llengua d'una única paraula, hi ha algunes mètriques que poden ser especialment útils per considerar.

En primer lloc, l'accuracy és una bona mètrica a considerar perquè mesura la proporció global de prediccions que són correctes. Això pot ser útil per obtenir una idea del rendiment global del model. No obstant, és important tenir en compte que l'accuracy sola pot ser enganyosa si les classes de les dades estan desequilibrades (és a dir, si hi ha un nombre desproporcionat de mostres pertanyents a una classe en comparació amb les altres). En aquest cas, altres mètriques com la precisió i el recordatori poden ser més informatives.

La mesura la proporció de prediccions veritables positives (és a dir, prediccions de llengua correctes) de totes les prediccions positives (és a dir, totes les prediccions per a una llengua determinada). El recall mesura la proporció de prediccions veritables positives de totes les mostres positives reals (és a dir, totes les mostres pertanyents a una llengua determinada). Aquestes mètriques poden ser útils per entendre el trade-off entre falsos positius i falsos negatius.

L'índex F1 és una mètrica que combina la precisió i el recall i sovint s'utilitza com una mesura resumida del rendiment d'un model de classificació. Es calcula com la mitjana harmònica de la precisió i el recall i és una bona mètrica per considerar quan voleu equilibrar la precisió i el recall .

Finalment, l'accuracy de top-3 és una mètrica que mesura la proporció de prediccions que són correctes i entre les 3 prediccions més probables. Això pot ser útil per entendre quant de segur és el model en les seves prediccions i com bé classifica les diferents opcions de llengua.

En resum, l'accuracy, la precisió, el recall, l'índex F1 i l'accuracy de top-3 són totes mètriques útils per considerar en avaluar el rendiment d'un model SVM per predir la llengua d'una única paraula. Aquestes mètriques poden ajudar a entendre el rendiment global del model, el tradeoff entre falsos positius i falsos negatius i la confiança i la classificació de les prediccions del model

Experiments realitzats

Com s'ha explicat anteriorment s'han creat tres conjunts, un d'entrenament, un de validació i un de test. Els percentatges per a cada conjunt s'han decidit fent unes proves amb els valors típics. Per al escalat de les dades s'han fet unes proves amb els diferents escaladors proporcionats per la llibreria scikit-learn:

- StandardScaler
- MinMaxScaler
- MaxAbsScaler
- RobustScaler
- Normalizer

Finalment s'ha decidit continuar amb el StandardScaler ja que s'ha considerat que és el més oportú per al problema i proporciona uns bons resultats.

Per a la reducció de la dimensionalitat només s'han fet proves amb la classe PCA a l'hora d'elegir el nombre de components s'han fet un conjunt de proves i finalment s'ha optat per mantenir una variança explicada del 99%, d'aquesta manera mantindrem la mayoria de la informació amb una quantitat menor de característiques. Per a aquesta variança se redueix el nombre de caractéristiques del conjunt {A,C,D} de 55 a 44.

Per als models s'ha decidit fer una prova amb els 4 *kernels* explicats anteriorment amb els seus hiperparametres per defecte.

Com ja s'ha explicat a l'apartat de característiques tenim 4 "grups" de característiques. Cap la possibilitat de que colcuna de les característiques creades siguin redundants, per averiguar quin conjunt de característiques és el millor s'executaran els 4 models, amb els hiper-paràmetres bàsics, amb totes les combinacions dels 4 grups de característiques.

A partir d'ara els grups de característiques tindran aquest nom: *Contament de caràcters* → A, *Contament de grups de caràcters* → B, *Prefixos i Sufixos* → C, *Llei de Zipf* → *D*.

Estudiant quines combinacions dels grups de característiques tenen millors resultats es podrà executar un GridSearchCV damunt dits conjunts, per obtenir els millors hiper-paràmetres. Cal recordar que aquesta cerca es farà emprant els conjunts d'entrenament i validació.

El Grid Search es farà amb els següents hiper-paràmetres:

- C: [0.1, 1, 5, 10]

- Gamma: [scale, auto, 1, 0.1, 0.01, 0.001]
- Kernel: Lineal o RBF
- Funció de decisió: 1 vs 1 o 1 vs la resta

Una vegada s'obtingui el millor conjunt de característiques i els millors hiper-paràmetres es podrà executar el model amb el conjunt de test, reservat fins el darrer moment.

Un experiment ideat es el de determinar l'idioma d'una oració sencera basant-se en l'idioma de les paraules individuals.

Resultats dels experiments

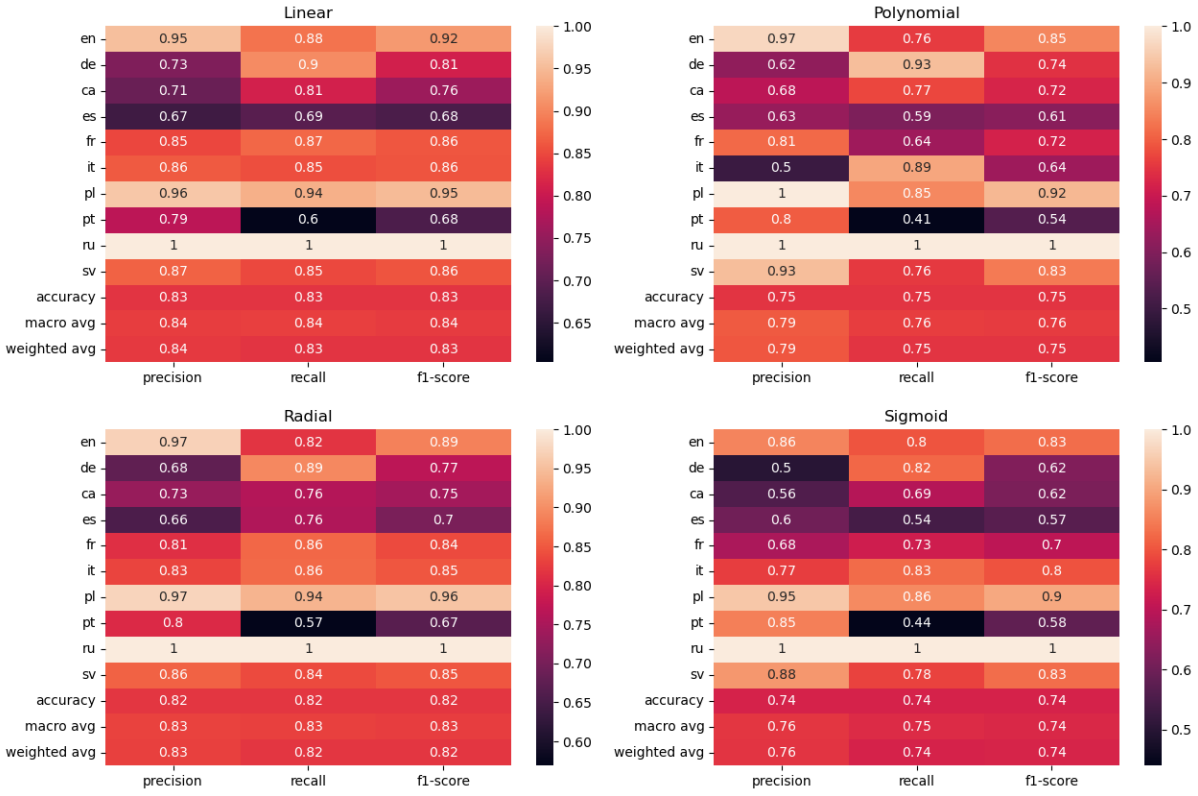
Models amb diferents Kernels

Per als 4 kernels mencionats anteriorment s'han obtingut els següents resultats utilitzant el dataset escalat amb el PCA aplicat:

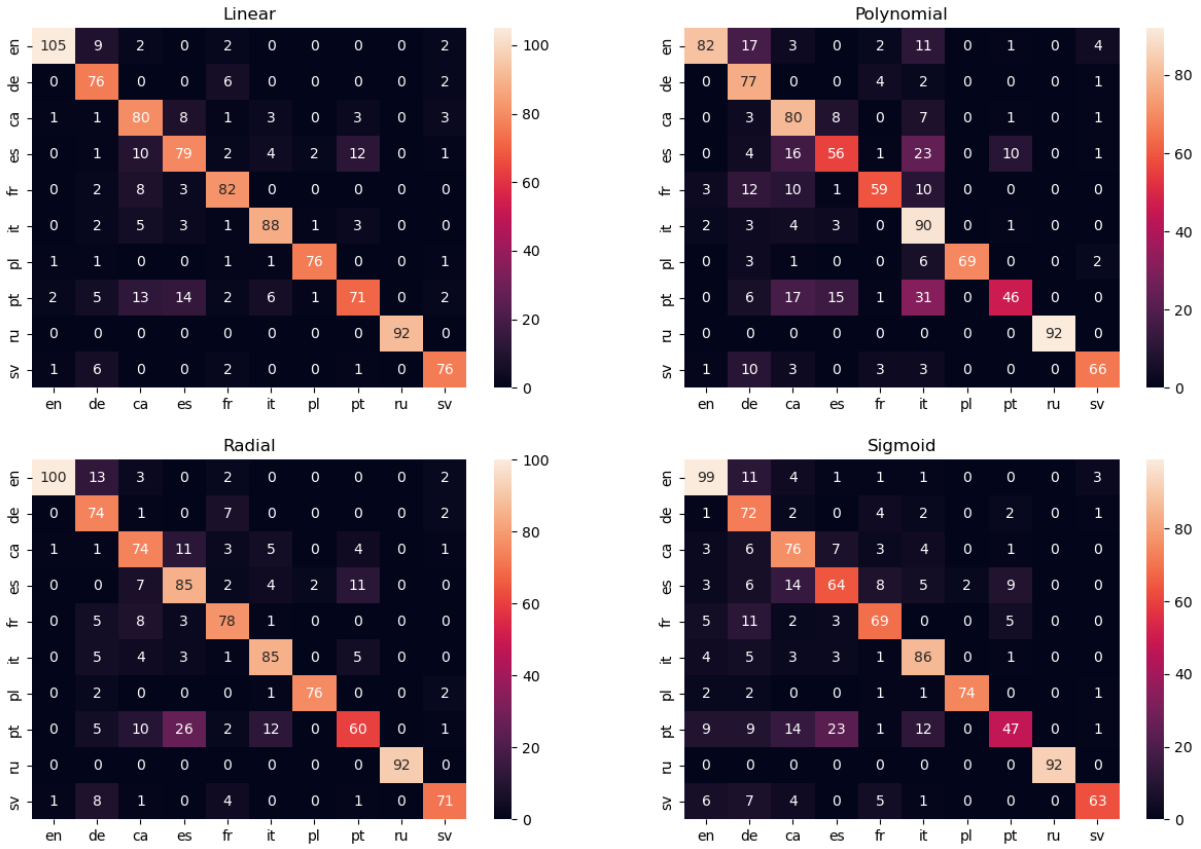
- **Accuracy:**
  Linear: 83.5%
  Polynomial: 72.57%
  Radial: 80.47%
  Sigmoid: 75.1%

- **Top 3 Accuracy:**
  Linear: 97.17%
  Polynomial: 95.75%
  Radial: 96.96%
  Sigmoid: 94.53%

- **Classification Report:**



- **Confusion Matrix:**



Com podem observar el model que obté els millors resultats es el que utilitza el kernel lineal, seguit del RBF. Tinguent en compte el rendiment el RBF tarda unes 10 vegades més que el lineal (0.4s vs 4s) a entrenar.

Millors conjunts de característiques

Així doncs tenim 15 possible combinacions dels conjunts de característiques:

Combinació	MODEL	Accuracy	Precision	F1-score	Top_3_score
{A}	Lineal	40.69%	50.55%	38.81%	66.19%
{B}	RBF	26.92%	30.83%	22.61%	53.34%
{C}	RBF	26.72%	40.51%	27.03%	56.28%
{D}	Poly	80.87%	81.04%	80.63%	96.46%
{A, B}	Lineal	45.95%	48.89%	45.15%	73.18%
{A, C}	RBF	45.75%	50.4%	44.69%	73.68%
{A, D}	Lineal	82.09%	82.3%	81.81%	96.56%
{B, C}	RBF	37.96%	37.91%	34.99%	63.66%

{B, D}	Lineal	81.28%	81.46%	80.94%	96.86%
{C, D}	Lineal	81.28%	81.63%	81.06%	96.86%
{A, B, C}	Lineal	49.49%	49.04%	48.02%	79.15%
{A, B, D}	Lineal	82.09%	82.14%	81.79%	96.76%
{A, C, D}	Lineal	82.19%	82.43%	81.93%	96.56%
{B, C, D}	Lineal	80.06%	80.14%	79.65%	95.65%
{A, B, C, D}	Lineal	81.48%	81.42%	81.13%	96.15%

Cal tenir en compte que els resultats poden variar entre execucions. És per això que no s'obtenen els mateixos resultats a n'aquest experiment que a l'anterior.

Com podem veure el conjunt D és de lluny el conjunt de característiques que ofereix una major explicabilitat de les dades.

Si es comparen els 15 conjunts es pot veure que el que té major F1-score és el subconjunt {A, C, D}, els següents experiments es faran amb aquests conjunts

Millors hiper-paràmetres

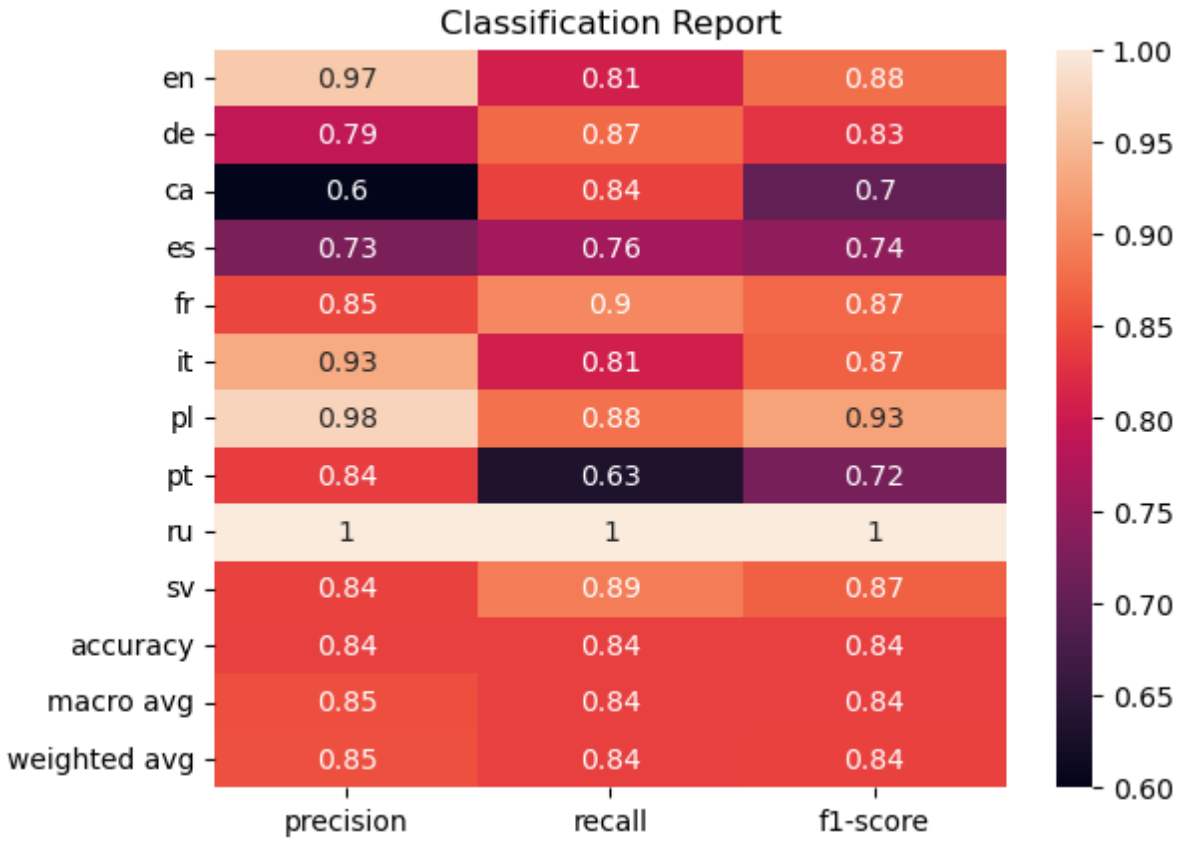
Una vegada trobat el millor conjunt de característiques, es pot executar la cerca dels millors paràmetres.

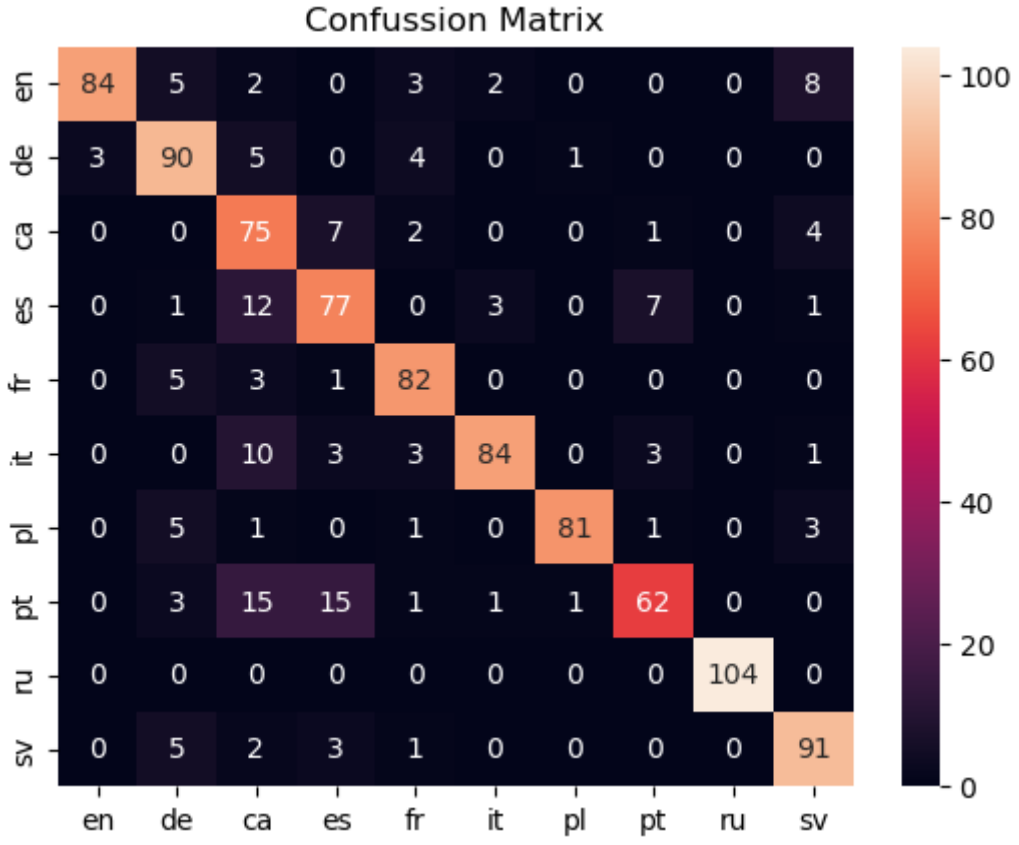
S'han obtingut els següents hiper paràmetres:

- C: 0.1
- Gamma: "scale"
- Kernel: Lineal
- Funció de decisió: 1 vs 1

Amb aquests hiper-paràmetres damunt el conjunt de test s'obtenen els següents resultats:

Acuracy	**Precision**	**F1-Score**	**Top 3 Accuracy**
84.01%	85.16%	84.06%	97.47%


Classification Report
en 0.97 0.81 0.88
de 0.79 0.87 0.83
ca 0.6 0.84 0.7
es 0.73 0.76 0.74
fr 0.85 0.9 0.87
it 0.93 0.81 0.87
pl 0.98 0.88 0.93
pt 0.84 0.63 0.72
ru 1 1 1
sv 0.84 0.89 0.87
accuracy 0.84 0.84 0.84
macro avg 0.85 0.84 0.84
weighted avg 0.85 0.84 0.84
precision
recall
f1-score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60


Confussion Matrix
en 84 5 2 0 3 2 0 0 0 8
de 3 90 5 0 4 0 1 0 0 0
ca 0 0 75 7 2 0 0 1 0 4
es 0 1 12 77 0 3 0 7 0 1
fr 0 5 3 1 82 0 0 0 0 0
it 0 0 10 3 3 84 0 3 0 1
pl 0 5 1 0 1 0 81 1 0 3
pt 0 3 15 15 1 1 1 62 0 0
ru 0 0 0 0 0 0 0 0 104 0
sv 0 5 2 3 1 0 0 0 0 91
en de ca es fr it pl pt ru sv
100
80
60
40
20
0

Predicció del llenguatge d'una frase

Per dur a terme aquest experiment s'han contabilitzat els idiomes predits de les paraules que formen una frase.

Frase d'exemple:
"Jugant a la baldufa amb els amics de classe em vaig fer mal al genoll dret"

Llenguatge predit:

ca: 44%
de: 31%
it: 12%

Conclusions

En general, el model SVM ha funcionat bé per determinar la llengua d'una paraula donada. El model ha estat capaç de classificar amb precisió la llengua de la majoria de les paraules de test, amb una precisió global del 84.06%. Això suggereix que el model SVM ha estat capaç d'aprendre eficaçment les característiques de cada llengua i utilitzar-les per fer prediccions.

Hi ha hagut alguns casos en què el model ha tingut dificultats per classificar amb precisió la llengua d'una paraula, especialment per a paraules que eren semblants en l'ortografia en idiomes parescuts com el portugués i el castellà. Això pot ser degut a una gran varietat de factors, com ara la mida limitada del conjunt de dades d'entrenament o la complexitat de les característiques de la llengua utilitzades pel model.

Tot i aquestes limitacions, el model SVM ha demostrat ser una eina prometedora per a la identificació de llengües. Amb una optimització ulterior i la inclusió de característiques de llengua addicionals, és probable que el rendiment del model pugui millorar encara més.

En general, aquest projecte d'aprenentatge automàtic ha demostrat l'eficàcia de l'ús de models SVM per a tasques d'identificació de llengües i ha posat de manifest la importància de seleccionar i enginyar amb cura les característiques per millorar el rendiment del model.