UNIVERSITAT DE BARCELONA

Creació d'un classificador de perfils segons el gènere de l'autor

Treball Final de Grau

Estudiant
David Cabestany Manen
d.cabestany@ub.edu

Tutora
Maria Taulé Delor
mtaule@ub.edu

Curs acadèmic 2020-2021

DECLARACIÓ D'AUTORIA

Amb aquest escrit declaro que sóc l'autor/autora original d'aquest treball i que no he emprat per a la seva elaboració cap altra font, incloses fonts d'Internet i altres mitjans electrònics, a part de les indicades. En el treball he assenyalat com a tals totes les citacions, literals o de contingut, que procedeixen d'altres obres. Tinc coneixement que d'altra manera, i segons el que s'indica a l'article 18, del capítol 5 de les Normes reguladores de l'avaluació i de la qualificació dels aprenentatges de la UB, l'avaluació comporta la qualificació de "Suspens".

Barcelona, a 17 de juny de 2021

David Cabestany Manen

Signatura:

Agraïments

Agraeixo als professors que he tingut al llarg de la carrera per tot el que m'han ensenyat, i tot el que he après d'ells, en especial a la meva tutora de TFG Mariona Taulé per haver-me guiat en l'elaboració d'aquest treball i per ensenyar-me inconscientment la constància i dedicació que necessita un projecte lingüístic; a la Montse Nofre per la seva inesgotable motivació, humilitat i ganes de superació; i a les dues per haver-me acceptat a l'STeL, que ha facilitat infinitament el camí a l'especialització en les Tecnologies del Llenguatge.

Voldria agrair al meu company Alejandro Ariza el temps que ha invertit a donar-me un cop de mà a l'hora d'entendre alguns algoritmes i també per donar-me pistes per tal de poder solucionar els entrebancs de la programació per mi mateix.

A ma mare vull agrair tot l'amor incondicional i els recursos que ha invertit en la meva educació al llarg de la vida i el seu inavaluable esforç que ha fet aguantant les meves dissertacions lingüístiques. A la família i amics, tot el suport moral que m'han ofert quan les coses no han anat com esperava; i per últim, vull agrair a la Cristina Peche les llargues converses sobre lingüística, antropologia, i filosofia en general, però sobretot per haver-me mostrat la teoria sobre el constructivisme identitari i cultural, que va suposar, en gran part, la inspiració d'aquest treball.

Índex

Índex de figures

Resum

En el següent estudi s'ha creat i posat a prova un model computacional per identificar el gènere dels autors de textos d'una manera automàtica, basant-nos en els textos del corpus CLEF2013, que han estat extrets de diferents blogs d'internet i dels seus comentaris. L'objectiu d'aquest estudi és elaborar un model de predicció del gènere d'autors desconeguts a partir del corpus emprat en la competició que es va dur a terme el 2013 a la PAN al CLEF. Avaluem dos models d'Aprenentatge Automàtic Supervisat per veure en quin d'ells obtenim millors resultats, la Regressió Logística i el Random Forest. El que millor funciona és el de Regressió Logística amb un encert de 0.64 punts.

Paraules clau: gènere, creació de perfils, identificació de trets, Aprenentatge Automàtic, Random Forest, Regressió Logística

En el siguiente estudio se ha creado y puesto a prueba un modelo computacional para identificar el género de los autores de textos de una forma automática, basándonos los textos del corpus CLEF2013 que han sido extraídos de diferentes blogs de internet y sus comentarios. El objetivo de este estudio es elaborar un modelo de predicción del género de autores desconocidos a partir del corpus empleado en la competición que se llevó a cabo en 2013 en la PAN al CLEF. Evaluamos dos modelos de Aprendizaje Automático Supervisado para ver en cuál de ellos obtenemos mejores resultados, la Regresión Logística y el Random Forest. El que mejor funciona es el de Regresión Logística con un acierto de 0.64 puntos.

Palabras clave: género, creación de perfiles, identificación de rasgos, Aprendizaje Automático, Random Forest, Regresión Logística

In the following study, we have created and tested a computational model to identify the gender of the authors of texts automatically. Based on the texts of the CLEF2013 corpus, extracted from different internet blogs and their comments. The objective of this study is to develop a model to predict the gender of unknown authors. For that, we used the corpus used in the competition that took place in 2013 at the PAN at CLEF. We evaluated two Supervised Machine Learning models to see which of them obtain better results, the Logistic Regression and the Random Forest. What works best is the Logistic Regression with a hit of 0.64 points.

Keywords: genre, profiling, feature identification, Machine Learning, Random Forest, Logistic Regression

1 INTRODUCCIÓ

1.1 *Marc de treball*

Els últims anys hem viscut un creixement en l'ús de recursos en línia que ha generat un enorme impacte en el nostre dia a dia. Recursos multimèdia com les xarxes socials, els blogs, les transmissions de vídeo en directe, els *mass media* o mitjans com el correu electrònic, ens faciliten una comunicació cada vegada més ràpida, gràcies, entre altres coses, als avenços tecnològics com l'accés a Internet.

La creació automàtica de perfils ha anat prenent importància a causa del gran buit informatiu sobre l'autoria dels continguts als quals tenim accés. Per aquest motiu s'està creant una necessitat de disposar d'eines capaces de detectar qui s'amaga darrere de determinats escrits i continguts (Rangel, Rosso, Potthast, Stein, & Daelemans, 2015). Dins del Processament del Llenguatge Natural (PLN) hi ha dues tasques concretes relacionades amb aquest objectiu: la creació automàtica de perfils (Author Profiling) i l'atribució d'autoria (Authorship Attribution). La creació automàtica de perfils utilitza les diferències lingüístiques que presenten els diferents textos per distingir classes d'autors en lloc d'autors individuals, a diferència de l'atribució d'autoria, que s'encarrega de determinar l'autor d'un text entre diferents conjunts tancats de candidats.

El present treball s'emmarca en l'àrea del PLN, i en concret en una de les seves tasques, la creació automàtica de perfils. El PLN i la Lingüística Computacional ens ofereixen la possibilitat d'emprar una metodologia basada en l'aprenentatge automàtic (Machine Learning - ML) per a l'elaboració d'un classificador automàtic de perfils, que identifiqui el gènere de la persona que ha escrit un determinat missatge. Per tal de dur-ho a terme utilitzem el corpus CLEF2013, que inclou comentaris extrets de diferents blogs. Aquest corpus ens servirà per entrenar i per avaluar dos models supervisats de ML, la Regressió Logística i els Random Forest per classificar automàticament el gènere dels autors dels comentaris (Home/Dona).

El nostre treball té una naturalesa interdisciplinària, ja que també partim del coneixement que ens ofereix la Sociolingüística i l'Anàlisi del Discurs. Tenim en compte i assumim els principis de la visió constructivista de l'adquisició i l'ús del llenguatge, que afirma que la llengua s'utilitza de manera diferent segons unes característiques socioculturals i psicològiques concretes adquirides al llarg del desenvolupament de l'individu (Córdoba, 2020; Vygotski, Kozulin, & Abadía, 1995). Per tant, hem creat un model computacional per veure si, realment, és possible classificar de manera automàtica els textos en relació amb el gènere dels autors i veure si l'eficiència del model és rellevant.

Des del punt de vista sociolingüístic, sempre ha suscitat interès poder interpretar els contrastos lingüístics que existeixen en els actes comunicatius dels diversos grups socials, grups de diferents zones geogràfiques, diferències d'edat, grups amb diferents llengües maternes,

distincions per sexe, gènere, diferents personalitats o, fins i tot, diferències psicològiques entre grups.

La variabilitat discursiva entre gèneres, es a dir, les diferències lingüístiques i comunicatives entre persones de diferents gèneres s'anomena *generolecte* i a les diferències que fan referència al sexe se'n diu *sexolectes* (Bedós, 2020). El sexe i el gènere no són el mateix, això hem de tenir-ho en compte a l'hora de fer una anàlisi lingüística. Entenem el sexe com una característica biològica amb la qual es neix. Entenem el gènere com la dimensió sociocultural al qual un individu es veu exposat en socialitzar-se. Dins dels *generolectes* trobem dues grans distincions, el *feminolecte* i el *masculinolecte* (Bedós, 2020). Probablement, avui en dia trobaríem necessària una distinció més fina i que tingui més en compte la varietat d'espectre del gènere, més adient amb la teoria dels rols de gènere, però el nostre corpus no descriu aquesta variabilitat i ens centrem en la classificació que hem descrit anteriorment.

Per tant deduïm que una variabilitat basada estrictament en el sexe no és viable, ja que totes les persones tenen les mateixes capacitats neurofisiològiques, independentment del sexe, però sí que existeix una variabilitat lingüística basada en les diferents assignacions socioculturals per a homes i dones dins d'una mateixa comunitat (Bedós, 2020).

Les diferències discursives entre diferents gèneres han estat àmpliament discutides i s'ha mantingut la idea que hi ha diferències consistents en els diferents discursos Asseverar quines són aquestes diferències és una mica més delicat. Hi ha molts treballs relacionats amb les diferències de tòpic entre els discursos d'homes i dones (Lakoff, 1973; Jones, 1980; Lozano Domingo, 1995; Mouton, 2003), normalment en aquestes afirmacions subjauen estereotips sobre el sexe que no esmentarem ni avaluarem en el nostre treball. Tot i que en estudis més actuals hi ha una tendència a relacionar les diferències a fenòmens d'educació i rols socials (Bedós, 2020; Mouton, 2003) el més important és que aquestes diferències, si hi són, sembla que s'esvaeixen quan entrem en registres més formals.

Els models que utilitzem per classificar els textos estan basats en freqüències de paraules. Autors com Argamon et al. (2009); Pennebaker et al. (2003) i Chambers and Schilling (2018) proposen utilitzar altres trets lingüístics que serveixen per distingir textos escrits per homes i dones, com per exemple, l'ús dels pronoms personals, l'ús d'atributs, de nominalitzacions, l'ús més freqüent de paraules específiques o de determinades estructures sintàctiques. En un futur ens agradaria millorar el nostre classificador incorporant-hi els trets que proposen aquests autors, que per falta de temps no ens ha estat possible.

En treballs realitzats en llengua anglesa, per exemple, s'ha trobat que hi ha una consistència en les freqüències dels usos de les formes morfològiques per part de cada parlant independentment del context en el qual s'emprin (Pennebaker et al., 2003). Aquests autors observen que les freqüències morfològiques són particulars de cada parlant, independentment del registre i del tòpic discursiu. També s'han trobat diferències en la manera que es presenten les entitats en el discurs i com s'interrelacionen amb l'emissor, és a dir, diferències en l'ús de pronoms i díctics segons la percepció del context per part del parlant, i que canvien segons l'adquisició lingüística

de l'individu.

Disposar de classificadros automàtics que identiqui si un text ha estat escrit per un home o una dona, actualment, poden ser de gran interès per a la recerca forense, la ciberseguretat i els estudis de màrqueting, també per a la moderació de continguts en línia de les xarxes socials com Twitter o Instagram.

Aquest treball s'estructura en 5 seccions: a la Secció 2 presentem els objectius i la de partida del treball; a la Secció 3 fem una breu descripció de l'estat de la qüestió; la Secció 4 tracta sobre la metodologia que hem fet servir; la Secció 5 presentem els resultats obtinguts amb els dos classificadors i, per últim, la Secció 6 exposem les conclusions del nostre projecte i les línies futures d'investigació.

2 OBJECTIUS I HIPÒTESI

A continuació exposarem els dos objectius principals i la hipòtesi del nostre treball.

- Objectiu 1: Crear un model basat en ML supervisat per classificar automàticament missatges segons si els ha escrit un home o una dona.
- Objectiu 2: Aprendre i aplicar la metodologia, els coneixements sobre aprenentatge automàtic, estadística i algoritmes necessaris per dur a terme el model de classificació i fer una recerca d'aquest tipus, així com ampliar els coneixements de programació adquirits en l'assignatura de Lingüística Computacional. En concret, millorar els meus coneixements de llenguatge de programació Python per implementar els dos models de classificació.

  Per tal de dur a terme aquests classificadors, partim de la següent hipòtesi:

- Les característiques de gènere mostren trets lingüístics diferents que ens permeten classificar textos segons el gènere dels autors, i per això utilitzem un corpus en què cada text s'ha anotat segons si els autors són homes o dones.

3 ESTAT DE L'ART

Anualment, des del 2013 al 2019 s'han celebrat unes competicions públiques amb diferents tasques (*shared tasks*) a la *PAN 2019 evaluation lab*, organitzades per la CLEF (Conference and Labs of the Evaluation Forum), que fan referència a l'anàlisi d'autoria i a la creació de perfils d'autors segons la forma d'escriure basades en els trets de gènere, edat o varietats dialectals, entre d'altres.

La iniciativa CLEF és un organisme auto-organitzat que té com a missió principal promoure la investigació, la innovació i el desenvolupament de sistemes d'accés a la informació amb èmfasi en la informació multilingüe i multi-modal. Explico més detalladament la competició del 2013 perquè utilitzem el corpus que es va proporcionar en aquella edició per entrenar els nostres models i els avaluo amb les mètriques utilitzades en la mateixa competició i els comparem.

En la competició de la PAN al CLEF del 2013 es van considerar els aspectes de gènere i edat com una part del problema de la detecció de perfils d'autor, focalitzat en l'anglès i el castellà. Fins aleshores els treballs de recerca en lingüística computacional i en psicologia social s'havien dut a terme principalment per a anglès. Aquesta tasca va ser la primera a dur a terme la detecció de perfils de manera automàtica en una llengua que no fos l'anglès, i va ser el castellà.

Segons Rangel et al. (2013), els concursants de la PAN al CLEF del 2013 es van centrar en utilitzar diferents tècniques per a la classificació dels textos segons el gènere de l'autor. Tots van utilitzar classificadors basats en models d'aprenentatge automàtic tradicionals i les diferències entre els diferents classificadors es centren en els trets que utilitzen els classificadors, en les tècniques de processament d'aquests texts i en els models de classificació.

En la fase de pre-processament fan servir *data cleaning* (Patra et al., 2013), (Moreau & Vogel, 2013), (Meina et al., 2013), (D. Weren et al., 2013), (Pavan et al., 2013), i, en canvi, (Lim et al., 2013) fa servir *principal component analysis* (PCA). Pel que fa als *features* (Lim et al., 2013), (Cruz et al., 2013), (Pavan et al., 2013), (Patra et al., 2013), (De-Arteaga et al., 2013), (Meina et al., 2013), (Flekova & Gurevych, 2013), (Alemán et al., 2013), (Santosh et al., 2013) van tindre en compte les freqüències dels signes de puntuació, majúscules, i en canvi, (Lim et al., 2013), (Meina et al., 2013), (Alemán et al., 2013), (Cruz et al., 2013), (Santosh et al., 2013) van fer servir mètodes de PoS tagging. És interessant destacar també que (Santosh et al., 2013), (Sapkota et al., 2013), (Meina et al., 2013) van tenir en compte enllaços i imatges,(Alemán et al., 2013), (Hernández et al., 2013) van tenir en compte les emoticones i les van descartar posteriorment per què no van obtenir els resultats esperats.

Pel que fa a les tècniques de processament de *features*, (Sapkota et al., 2013), (Patra et al., 2013) (Lim et al., 2013), (Mechti et al., 2013), (Caurcel Díaz & Gómez Hidalgo, 2013), (Flekova & Gurevych, 2013), (Meina et al., 2013), (Cruz et al., 2013), (Santosh et al., 2013), (Pavan et al., 2013), (Hernández et al., 2013) van inclinar-se per escollir tècniques com *Latent Semantic Analysis* (LSA), *Bag of Words* (BoW), TF-IDF, *Entropy Based-Words*, entre d'altres. Pel que fa als models de ML de classificació (models de predicció), van utilitzar arbres de decisió (Santosh et al., 2013), (Patra et al., 2013), (Mechti et al., 2013), (Gillam, 2013) i (D. Weren et al., 2013), però no especifiquen de quin tipus. Van utilitzar *Support Vector Machines* (Lim et al., 2013), (Cruz et al., 2013) i (Sapkota et al., 2013). Logistic Regression el van utilitzar (De-Arteaga et al., 2013), (Flekova & Gurevych, 2013), *Naïve Bayes* (Meina et al., 2013), *Maximum Entropy* (Pavan et al., 2013), *Stochastic Gradient Descent* (SGD) va ser emprat per (Caurcel Díaz & Gómez Hidalgo, 2013), i per últim Random Forest va ser emprat per (Alemán et al., 2013).

Avui en dia, tot i que moltes d'aquestes tècniques es continuen utilitzant, hi ha una forta

tendència a utilitzar les anomenades tècniques de *Deep Learning* com les *Recurrent Neural Networks* (RNN). Dins de les RNN podem trobar la *Long-Short Term Memory* (LSTM) (Hochreiter & Schmidhuber, 1997), o la *Gated Recurrent Units* (GRU) (Cho et al., 2014). A partir de la competició PAN al CLEF del 2017 (Rangel, Rosso, Potthast, & Stein, 2017) es van començar a utilitzar sistemes basats en xarxes neuronals com les RNN, *Convolutional Neural Networks* (CNN) i *Deep Averaging Networks* (DAN) (Iyyer, Manjunatha, Boyd-Graber, & Daumé III, 2015).

Estudis recents en PLN han observat un bon comportament en l'ús de tècniques *context-dependents* i *task-free* en la comprensió del llenguatge natural (Polignano et al., 2019), també anomenades *transformers* que ha facilitat la construcció de models amb més capacitat inspirats en el pioner *Tensor2tensor* de Google (Wolf et al., 2020). Actualment hi ha una forta tendència a utilitzar *transformers* com ELMo (Embeddings from Language Models) (Peters et al., 2018), GPT-3 (Generative Pre-trained Transformer 3) (Brown et al., 2020), o *transformers* basats en BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), aquest últim també és propietat de Google.

Actualment, la detecció automàtica de perfils està més centrada en determinar si dos o més textos han estat escrits pel mateix autor. Per tant, seria una tasca a mig camí entre la detecció automàtica de perfils i l'atribució d'autoria, perquè es tracta de diferents textos "emmascarats", és a dir, en aquells en què es varia l'estil, el registre i diferents aspectes lingüístics de manera intencionada pels autors, per tal d'ocultar la seva autoria.

4 METODOLOGIA

En aquesta secció parlarem sobre la metodologia que hem emprat per elaborar els nostres models de detecció de gènere dels autors i els processos que hem tingut en compte per elaborar-los, així com les etapes que hem seguit per implementar cada element del nostre model (Figura 1).

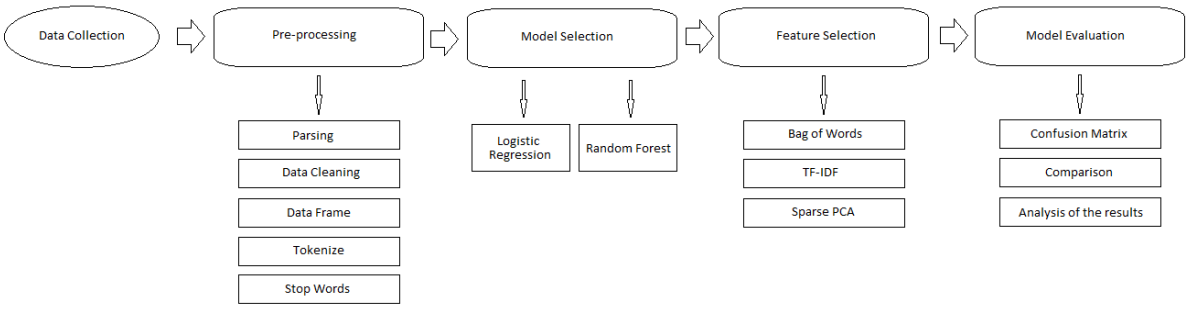


Figura 1: Pipeline per a la creació del model

Per tal de dur a terme els classificadors basats en ML, la metodologia aplicada s'estructura en cinc parts que es descriuen e les seccions següents :

Primer explicarem en què consisteix la tasca a la Secció 4.1. Després presentarem el corpus d'aprenentatge i avaluació en la Secció 4.2. En la Secció 4.3 descriurem els mètodes de pre-processament que hem utilitzat en el nostre classificador. A continuació, els models d'entrenament que utilitzarem per als classificadors a la Secció 4.4, després a la Secció 4.5 explicarem la sel·lecció dels trets o *features*. Per últim, parlarem de l'avaluació que hem dut a terme dels models.

4.1 *Definició de la tasca*

La tasca consisteix a crear un classificador binari que a partir d'un text pla que pren com a *input*, generi un *output* on se li assigni una etiqueta amb el valor (*'male'/'female'*).

id	text	target
bb049b7346	HOLAS ME LLAMO MANI MMM LA FOTO ESTA UN POCO TILDEADA PERO SE NOTA como soy y bueno si me quieren conocer supongo q saben q tienen q hacer pero les dejare mi msn bye soy bien cool segun otros/as	???
23g4j57s89	hoy esta lloviendo y no me gustan los dias asi, por que no me dejan hacer lo que mas me gusta.....jajajjajajaaj	???

Taula 1: Exemple d'input que els classificadors hauran de processar

En l'exemple de la Taula 1 tenim dos comentaris input que els classificadors hauran d'etiquetar. En el primer cas, l'etiqueta hauria de ser '*male*' i en el segon hauria de ser '*female*'.

4.2 *Corpus CLEF2013*

En aquesta secció descrivim el corpus CLEF2013 que es el que s'ha fet servir per entrenar i avaluar els dos models del classificador.

El corpus CLEF2013 està format per 75.900 comentaris en castellà i anotats segons el gènere de l'autor del comentari. Aquest corpus ens l'han proporcionat els organitzadors de la tasca PAN al CLEF del 2013.

A l'hora de recollir les dades per crear el corpus es van seguir dos criteris bàsics. Recollir temàtica diversa i no estar escrits per xat-bots (Rangel et al., 2013).

Els comentaris s'han extret del repositori públic i obert Netlog, que ja no és disponible des de l'abril de 2015, perquè contenia dades demogràfiques de l'autor, com el sexe i l'edat. Un cop seleccionats, es van agrupar els comentaris per gènere de l'autor, i només es van tenir en compte els autors amb més d'una publicació, ja que els autors que tinguessin poques publicacions, o bé, que fossin molt breus no podien encaixar en un marc d'avaluació realista.

Les publicacions d'aquells autors que tinguessin més de 1.000 paraules es van dividir en dos o més arxius diferents. Els comentaris d'un mateix autor només es van incloure o bé, en el corpus d'entrenament, o bé en el de avaluació per tal d'evitar deteccions errònies innecessàries.

Tal i com podem observar a la Taula 2, el CLEF2013 està dividit en dues parts o *datasets*. El dataset d'entrenament (CLEF2013_Train) amb 75.900 comentaris anotats i el dataset de test (CLEF2013_Test) amb 8.160 comentaris sense anotar. El corpus CLEF2013 conté en total 84.060 comentaris. Una part amb anotacions a la qual anomenem corpus d'entrenament, una part sense anotacions a la qual anomenem corpus de test i un arxiu de verificació, que inclou l'anotació del corpus de test per avaluar els sistemes.

Lang	Gender	Age	No. of Authors			
			Training	Test	$\sum Trainig$	$\sum Test$
es	male	10s	1250	144	39750	4080
		20s	21300	2304		
		30s	15400	1632		
	female	10s	1250	144	39750	4080
		20s	21300	2304		
		30s	15400	1632		
$\sum$			75900	8160	75900	8160

Taula 2: Anàlisi estadístic del corpus CLEF2013 (Rangel et al., 2013)

El corpus CLEF2013 està anotat en XML, la Figura 2 és una mostra de com està estructurat el corpus. Com podem observar, per cada comentari podem veure la llengua (lang="es") l'edat (age_group="20") i el gènere(gender="female"). En aquest treball, per entrenar el model només tindrem en compte l'etiqueta *gender*, ho expliquem amb més detall a la Secció 4.3.1.

```
1 <root>
2 <author lang="es" gender="female" age_group="20s">
3     <conversations count="1">
4       <conversation id="3cf2398d36dd2f13bed">
5             &gt; los amigos son como las estrellas estan hai on ke no
                  las  puedas ver./&gt;
6         </conversation>
7     </conversations>
8 </author>
9 </root>
```

Figura 2: Estructura Corpus CLEF2013

4.3 Pre-processing

El primer pas en tot estudi lingüístic és entendre la informació que tenim i què volem saber o analitzar. En lingüística computacional aquest pas cobra més importància, fins i tot, ja que hem d'entendre la informació que tenim, interpretar-la i fer-la-hi entendre a l'ordinador mitjançant el codi de programació. Hem de saber quins passos cal seguir per a analitzar un text, entendre els passos a seguir i perquè han de fer-se, després poder traduir-los a un llenguatge que sigui comprensible per a l'ordinador i així automatitzar el procés.

El preprocessament consisteix a preparar les dades perquè els classificadors puguin seleccionar els trets que tindran en compte a l'hora de classificar els textos que li proporcionem a l'*input*.

Un cop disposem del corpus, el següent pas és el preprocessament del corpus per tal que l'algoritme de processament pugui interpretar correctament les dades. En aquest treball hem aplicat 5 etapes diferents: el *Parsing* de les dades, el *Data Cleaning*, l'organització de les dades en *Data Frames*, el procès de *Tokenització* i, per últim, què són les *stop words*.

4.3.1 *Parsing*

En el segment del document XML que hem adjuntat a la Figura 2 podem observar quatre línies d'obertura(<root>, <author>, <conversations> i <conversation>) i tres línies de tancament de les dades (</root>, </author>, </conversations> i </conversation>). Aquestes etiquetes en les línies ens indiquen que tenim 4 elements principals en aquest document.

En aquest cas la informació que ens interessa és el segon atribut de la segona línia (gender = "female") i l'atribut de la tercera línia (id = "3cf2398d36dd2f13bed"), què és una sèrie alfanumèrica que serveix per identificar l'arxiu. El primer element correspon al nostre *target*

i el segon element és l'identificador del text, ha d'emmagatzemar-se per a poder comparar el resultat amb l'arxiu de verificació què és aquell document on hi ha les etiquetes reals dels arxius els quals prediem per tal de comparar els resultats, per tant, és important que l'id coincideixi amb l'índex de l'arxiu.

Després hem d'assignar el text a aquests dos atributs, és a dir, quan organitzem les dades, crearem un *DataFrame* on hi haurà una columna que contingui els textos, una l'atribut 'gender' i l'altra l'atribut 'id'. Cada fila correspondrà a un text i de manera adjacent trobarem el seu id i el seu gènere.

4.3.2 *Data Cleaning*

El dataset que utilitzem conté el text, el qual abans d'organitzar-lo en un *Data Frame*, li haurem d'aplicar el que anomenem *data cleaning*. Moltes anotacions i etiquetes no ens interessen perquè no les fem servir, per tant tota aquesta informació irrellevant per a nosaltres i que ens dificulti l'anàlisi dels textos l'haurem d'eliminar, com per exemple, les etiquetes que fan referència als salts de línia, sagnats i altres codificacions. Per tant, el *data cleaning* fa referència a l'eliminació del soroll dels textos del corpus CLEF2013.

4.3.3 *Data Frame (Pandas)*

Un cop tenim el text net procedim a estructurar les dades que conté. Per fer-ho utilitzem un *Data Frame* (des d'ara DF) de dues dimensions (Taula 3), és a dir, les dades s'organitzen en l'eix X (columnes) i en l'eix Y (files). Com es pot veure a la Taula 3 la informació s'estructura en tres columnes en què s'inclourà l'id, el text i el target (*'male'/'female'*), a més a més, les dades s'emmagatzemen etiquetades, això vol dir que sempre que no tallem la taula, totes les dades que hi hagi en una mateixa fila estaran vinculades tot i trobar-se organitzada en diferents columnes. Podem pensar en el DF com una mena de base de dades on podem emmagatzemar dades de tipus numèric i de tipus textual. Per organitzar les dades en un DF hem utilitzat les llibreries Pandas (Wes McKinney, 2010) i Numpy (Harris et al., 2020).

id	text	target
bb049b8093	lo esencial es invisible a los ojos [...]	female
cd52b2aad3	vamos a tratar de dar un contenido de utilidad [...]	male
aa7b78896a	Muy bien desde hoy empezare a manejar el blog [...]	male
3033ac5f3b	dejame mostrarte los dulces placeres del [...]	male
2c5b060cab	hola espero su firma y q las sigan pasando [...]	female
4061e9bbc6	En la vida hay que saber a quién sonreirle [...]	female

Taula 3: Extracte DF d'entrenament

El motiu pel qual utilitzem aquesta organització és perquè ens va bé que l'id del text estigui a l'esquerra i el '*target*' a la dreta. D'aquesta manera, a l'hora d'utilitzar el corpus de test del CLEF2013, no tindrem el '*target*' i, per tant, que quedi a la dreta ens ajuda a no haver de reestructurar les columnes del Data Frame test. A més, tot i que no és necessari estructurar-ho d'una manera, o bé d'una altra, ens ha semblat que fer-ho així era més intuïtiu. La Taula 3 exemplifica el DF d'entrenament i el DF d'avaluació tindrà el mateix format però amb el tret '*target*' buit per a la informació que el classificador haurà d'assignar automàticament un cop entrenat.

4.3.4 *Tokenitzar (NLTK)*

Tokenizar és separar el text en les unitats mínimes que el conformen. A les unitats resultants se les coneix com a *token*, a grans trets un token pot semblar una paraula, de fet, ho és depenent de què considerem paraula, però el token, en PLN pot englobar signes de puntuació, números, abreviatures i parts de paraula.

Per exemple, podríem dir que 'ull de bou' són tres paraules diferents, 'ull', 'de' i 'bou', ja que per separat tenen significat o una funció específica. Però mentre que 'ull de bou' podria conformar una sola paraula segons la concepció que tinguem, sempre estarà formada per tres tokens diferents, si volguéssim fer que 'ull de bou' fos un sol token hi ha maneres de fer-ho, per exemple escriure 'ull_de_bou', però és poc freqüent. Hi ha definicions de paraula que es corresponen a la de token, però d'aquesta manera és més senzill evitar confusions. Per tant, definim token com tot aquell contingut textual i numèric que hi ha entre dos espais, o en tot cas, salts de línia o puntuació, dins d'un text.

Els tokens de cada arxiu els afegim en una nova columna del Data Frame per tal de poder accedir-hi més endavant. Ja que en fer el TF-IDF els necessitarem. Al nostre Data Frame li hem afegit aquesta columna, on hi tenim els tokens de cada text (Taula 4).

id	tokens
bb049b8093	'lo', 'esencial', 'es', ' invisible', 'a', 'los', 'ojos', '!', '!', '!', 'recuerdenlo',
cd52b2aad3	'vamos', 'a', 'tratar', 'de', 'dar', 'un', 'contenido', 'de', 'utilidad',
aa7b78896a	'Muy', 'bien', 'desde', 'hoy', 'empezare', 'a', 'manejar', 'el', 'blog',
3033ac5f3b	'dejame', 'mostrarte', 'los', 'dulces', 'placeres', 'del', 'infierno', 'sin',
2c5b060cab	'hola', 'espero', 'su', 'firma', 'y', 'q', 'las', 'sigan', 'pasando', 'bonito'
4061e9bbc6	'En', 'la', 'vida', 'hay', 'que', 'saber', 'a', 'quién', 'sonreirle',

Taula 4: Columna DF amb els tokens

4.3.5 *Stop Words (Scikit-learn)*

Finalment, les *stop words* són totes aquelles paraules que no aporten riquesa semàntica al text, com per exemple les preposicions, conjuncions, articles, pronoms, connectors, així com els verbs *'ser', 'estar', 'haber'* i *'tener'* quan tenen funció auxiliar.

Queda a criteri del lingüista decidir com interpretar aquestes paraules, hi ha casos en els quals pot ser útil tenir-les en compte i n'hi ha que no són rellevants. Habitualment s'opta per no interpretar-les ni tenir-les en compte o, sí més no, aïllar-les en el cas de voler tenir-les en compte més endavant com a part de l'anàlisi.

En un principi, en el nostre treball vam decidir no tenir en compte les *stop words*, ja que en basar la nostra anàlisi en freqüències de paraules, ho podíem gestionar directament amb el mètode estadístic TF-IDF, perquè treu rellevància als tokens amb una freqüència d'aparició més elevada en comparació als demés tokens, tal com expliquem en l'apartat 4.5.2.

Tot i així, per tal d'analitzar la importància de les *stop words* en la classificació de gènere, hem fet l'anàlisi TF-IDF tenint en compte les *stop words* i sense tenir-les en compte, tot i que en principi, com expliquem més endavant no és realment necessari perquè el TF-IDF ja treu importància a les paraules que es repeteixen. La Taula 5 inclou la llista completa amb les 308 paraules que hem tingut en compte com stop words.

stop words
'a','al','algo','algunas','algunos','ante','antes','como','con','contra','cual','cuando', 'de','del','desde','donde','durante','e','el','ella','ellas','ellos','en','entre','era', 'erais','eran','eras','eres','es','esa','esas','ese','eso','esos','esta','estaba','estabais', 'estaban','estabas','estad','estada','estadas','estado','estados','estamos','estando','estar', 'estaremos','estará','estarán','estarás','estaré','estaréis','estaría','estaríais','estaríamos' ,'estarían','estarías','estas','este','estemos','esto','estos','estoy','estuve','estuviera', 'estuvierais','estuvieran','estuvieras','estuvieron','estuviese','estuvieseis','estuviesen', 'estuvieses','estuvimos','estuviste','estuvisteis','estuviéramos','estuviésemos','estuvo', 'está','estábamos','estáis','están','estás','esté','estéis','estén','estés','fue','fuera', 'fuerais','fueran','fueras','fueron','fuese','fueseis','fuesen','fueses','fui','fuimos', 'fuiste','fuisteis','fuéramos','fuésemos','ha','habida','habidas','habido','habidos','habiendo', 'habremos','habrá','habrán','habrás','habré','habréis','habría','habríais','habríamos', 'habrían','habrías','habéis','había','habíais','habíamos','habían','habías','han','has','hasta', 'hay','haya','hayamos','hayan','hayas','hayáis','he','hemos','hube','hubiera','hubierais', 'hubieran','hubieras','hubieron','hubiese','hubieseis','hubiesen','hubieses','hubimos', 'hubiste','hubisteis','hubiéramos','hubiésemos','hubo','la','las','le','les','lo','los','me','mi', 'mis','mucho','muchos','muy','más','mí','mía','mías','mío','míos','nada','ni','no','nos', 'nosotras','nosotros','nuestra','nuestras','nuestro','nuestros','o','os','otra','otras','otro', 'otros','para','pero','poco','por','porque','que','quien','quienes','qué','se','sea','seamos', 'sean','seas','seremos','será','serán','serás','seré','seréis','sería','seríais','seríamos', 'serían','serías','seáis','sido','siendo','sin','sobre','sois','somos','son','soy','su','sus', 'suya','suyas','suyo','suyos','sí','también','tanto','te','tendremos','tendrá','tendrán', 'tendrás','tendré','tendréis','tendría','tendríais','tendríamos','tendrían','tendrías','tened', 'tenemos','tenga','tengamos','tengan','tengas','tengo','tengáis','tenida','tenidas','tenido', 'tenidos','teniendo','tenéis','tenía','teníais','teníamos','tenían','tenías','ti','tiene', 'tienen','tienes','todo','todos','tu','tus','tuve','tuviera','tuvierais','tuvieran','tuvieras', 'tuvieron','tuviese','tuvieseis','tuviesen','tuvieses','tuvimos','tuviste','tuvisteis', 'tuviéramos','tuviésemos','tuvo','tuya','tuyas','tuyo','tuyos','tú','un','una','uno','unos', 'vosotras','vosotros','vuestra','vuestras','vuestro','vuestros','y','ya','yo','él','éramos'

Taula 5: Llista d'*stop words*

Ens agradaria fer notar que aquesta llista d'*stop words* l'hem extret del repositori de Savand, Alireza a (*Python Package Index - PyPI*, 2003), que es va fer el 2018. Hem notat que li falta alguna revisió, ja que s'hi ha d'afegir algunes formes com l'infinitiu del verb ser, entre d'altres. Nosaltres no hem modificat el diccionari en el nostre classificador.

4.4 Selecció del model d'entrenament

En aquest apartat parlarem dels diferents algoritmes que hem utilitzat per fer les prediccions sobre el gènere dels autors del CLEF2013_Test. Hem utilitzat dos models d'entrenament estadístics per a la tasca de la classificació binària, el model de Regressió Logística de la biblioteca Scikit-learn i el model de Random Forest, també de la mateixa biblioteca.

4.4.1 *Logistic Regression*

La Regressió Logística és un model estadístic emprat en la predicció d'esdeveniments categòrics, és a dir, variables que puguin adoptar un nombre definit de categories, a partir d'un nombre indeterminat de variables independents en les quals es basa per crear generalitzacions que farà servir per crear les prediccions. En el nostre cas com només tenim dues categories, podem fer una Regressió Logística unidimensional, perquè les variables només poden agafar el valor 0 que correspondria a l'atribut '*male*'; o bé el valor 1, que correspondria a l'atribut '*female*' del CLEF2013.

4.4.2 *Random Forest*

Per entendre què és un Random Forest, primer hem d'entendre què és un arbre de decisió. Un arbre de decisió és un tipus d'algoritme que s'organitza en forma d'arbre i que conté claus amb condicions per tal de poder prendre decisions de predicció, gràficament recorda molt a un diagrama de flux.

A la Figura 3 podem observar com funciona un arbre de decisió, en aquest cas es tracta d'un classificador de flors d'Iris segons la seva taxonomia. La Figura està extreta de la documentació de scikit-learn, Secció 1.10 (Pedregosa et al., 2011).

És fàcil que un arbre de decisió dissenyat manualment *ad hoc* acabi essent massa gran (overfitting), és a dir que tingui massa condicions, si volem que tingui en compte tants trets com creguem necessaris per a una classificació i acabi per ocasionar que el model de ML no pugui fer prediccions suficientment generals per classificar dades noves en grups significants. També pot succeir que a l'hora de dissenyar l'arbre ens descuidem trets que resultin ser representatius i ens quedi un arbre massa petit (underfitting), que pot ocasionar unes prediccions massa generals i la classificació de les dades entrin en un sol grup. En el nostre cas, un *underfitting* ocasionaria que la predicció ens classifiqui tots els nous textos en un sol gènere, ja sigui a la categoria '*male*' com a la '*female*'.

El model Random Forest el que fa és crear un nombre determinat d'arbres de decisió, on cada arbre contindrà un nombre aleatori de condicions. Un cop creats *n* arbres es fa una votació

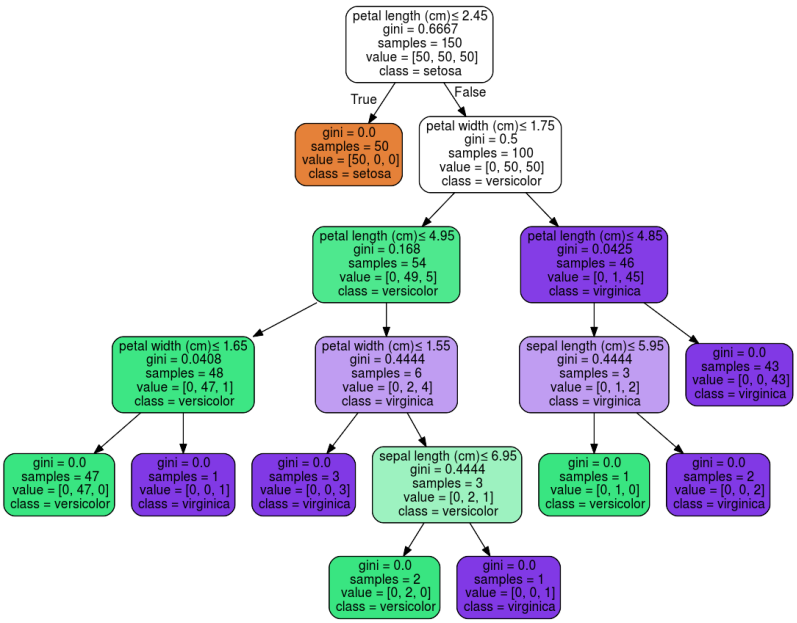


Figura 3: Arbre de decisió del classificador d'Iris

per veure quins arbres són els que s'adiuen millor per a cada text, i a partir d'aquest arbre es classifica el text segons la probabilitat de ser '*male*'/'*female*'.

En el nostre classificador, a l'hora d'estipular el nombre d'arbres que necessitaríem, vàrem fer tres proves, una primera prova amb 2 arbres (Figura 4b), una segona prova amb 100 (Figura 4c) i una última amb 800 arbres (Figura 4d). La diferència dels resultats obtinguts amb la primera prova i la segona han estat significatius, mentre que les diferències entre la segona prova i l'última, tot i haver-hi una quantitat d'arbres molt superior, no han estat gaire significatives qualitativament, però el temps de processament sí que ha estat notablement superior, de fet, la proporció ha estat molt més lenta la tercera respecte a la segona que la segona respecte de la primera.

4.5 Selecció dels trets (*features*)

La següent etapa consisteix en la relació de trets, és a dir, seleccionar quins trets voldrem que el classificador tingui en compte per classificar.

La selecció de trets, en el nostre projecte, consta de 3 parts, el *Bag of Words* (BoW) (Secció 4.5.1), la vectorització (Secció 4.5.2), i l'Sparse PCA (Principal Component Analysis) (Secció 4.5.3). Per a la vectorització, hem escollit el mètode TF-IDF (Term Frequency – Inverse Document Frequency). Hem de tenir en compte que nosaltres volem explicar el mètode Bag of Words per tal que s'entengui el seu funcionament, ja que tècnicament l'utilitzem, però el nostre classificador fa el BoW automàticament quan apliquem el TF-IDF, ja que la llibreria que

emprem no requereix que ho fem per separat.

4.5.1 *Bag of Words*

Bag of words (BoW) és el procés en el qual es crea una matriu, o una taula, on cada token correspon a una columna i cada fila emmagatzema el número d'aparicions de cada token en cada text (Zhang, Jin, & Zhou, 2010). Cada fila queda enllaçada a un text amb un id diferent. A continuació mostrem una recreació de com funcionaria el nostre BoW (Taula 6).

id	target	lo	esencial	es	invisible	a	los	ojos	!
bb049b8093	female	1	1	1	1	1	1	1	3
cd52b2aad3	male	1	0	0	0	1	1	0	0
aa7b78896a	male	0	0	1	0	1	0	0	0
3033ac5f3b	male	0	0	0	0	0	1	0	0
2c5b060cab	female	0	0	0	0	0	0	0	0
4061e9bbc6	female	0	0	0	0	1	0	0	0

Taula 6: Recreació del Bag of Words del CLEF2013

Com es pot observar, el que fa el BoW és convertir cada token que apareix en cada arxiu i el transforma en noves columnes, sempre que no hagi aparegut anteriorment, és a dir, ens quedaria tantes columnes com types hi hagi en el corpus CLEF2013_Train. Conservem la columna de l'id que correspon a l'arxiu i també la columna 'target', d'aquesta manera no perdem les referències. La matriu es va omplint amb la quantitat de cops que apareix en l'arxiu el token que encapçala la columna. Si prenem de referència la Taula 3 veurem que la Taula 6 va enmagatzemant el nombre de cops que apareix els tokens que hem llistat.

4.5.2 *Vectorització*

S'anomena vectorització el procés en el qual se li assigna un valor numèric a cada token a partir de les dades que hem obtingut en la matriu del BoW, aquest valor numèric l'obtenim aplicant unes fórmules matemàtiques. Els valors numèrics que assignem als tokens descriuen la representativitat de cada token en relació amb les freqüències relatives dels tokens en el còmput global de textos (*'male'/'female'*) del corpus.

Les fórmules matemàtiques que hem comentat són d'un model d'anàlisi estadística anomenada TF-IDF, és relativament senzill d'entendre i senzill d'implementar en un codi de Python. El TF-IDF és el producte de dos càlculs: freqüència dels termes (TF) i la freqüència inversa dels documents (IDF).

En el cas de la freqüència dels termes TF(t,d) la forma més senzilla de determinar el seu

valor és calcular el nombre de vegades que apareix el terme 't' en el document 'd'. Si denotem la freqüència de 't' per a F(t,d), llavors l'esquema 'TF' simple és TF(t,d) = F(t,d).

Una manera més eficient de calcular la 'TF' és dividint la freqüència del terme que busquem per la freqüència augmentada (freqüència del terme que apareix més vegades en el text) (Ramos et al., 2003)(Wikipedia, 2013). Això s'utilitza com un factor de ponderació en la recuperació d'informació. El valor TF-IDF augmenta de manera inversament proporcional a la rellevància de la paraula en un corpus. Així tenim que una paraula amb un TF-IDF de 2.03 és menys rellevant que una paraula amb un TF-IDF de 0.34, i per tant no la tindrem tant en compte com la segona.

4.5.3 *Sparse PCA*

Quan tenim tota una matriu de vectors després de fer el TF-IDF, ens trobem que la matriu és excessivament extensa donat que cada coordenada equival a un token. Pel mateix motiu obtenim una matriu amb molts espais buits, ja que no tots els tokens apareixen en tots els documents. Per tant, ens trobem en una situació en què la dimensió de la matriu que conté tots els vectors d'entrenament del model és desproporcionadament gran.

El mètode Sparse PCA (Sparse Principal Component Analysis) el que fa és fer un càlcul matemàtic que redimensiona una matriu amb la menor pèrdua d'informació possible, per tal que no hi hagi espais en buit i que el model d'entrenament requereixi menys recursos. S'ha d'anar amb compte, perquè si redimensionem la matriu en excés, podríem tindre una pèrdua que afectés els resultats de la classificació. Nosaltres fem servir el que ens proporciona la llibreria d'scikit-learn, que funciona segons la funció següent (Pedregosa et al., 2011):

$$(U^*, V^*) = \underset{U,V}{\arg\min} \frac{1}{2}||X - UV||_2^2 + \alpha||V||_1$$
$$\text{subject to } ||U_k||_2 = 1 \text{ for all } 0 \leq k < n_{components}$$

4.6 *Avaluació del model*

Un cop entrenat el nostre classificador amb els models de ML, hem de veure quina puntuació obtenen, és a dir, quins resultats dona. Aquesta puntuació s'anomena *score* i és un valor d'entre zero i u. Un cop coneixem aquest percentatge podem començar a fer conjectures de quins poden ser els possibles motius pels quals el model hagi obtingut una puntuació o una altra.

Normalment el model s'entrena primer amb una part del corpus d'entrenament que es reserva especialment per fer el que s'anomena un *Early Bird*, és a dir, una primera presa de contacte amb els resultats de la predicció, per tal de veure si el model funciona com s'espera,

abans d'analitzar el CLEF2013_Test. Nosaltres vam fer el *Early Bird* només amb la Regressió Logística amb un terç del Data Set d'entrenament del CLEF2013, ja que va ser el primer model que vam implementar i volíem estar segurs que tot el que havíem fet fins al moment funcionava com esperàvem, amb el Random Forest no va caler fer-ho, ja que a diferència dels participants de la PAN al CLEF, nosaltres disposàvem del CLEF2013_Test per poder fer les proves directament. L'Early Bird que vam fer ens va retornar una *score* de 0.879, amb una *score* d'aquestes característiques podem dir que el model funciona dins del previst, tot i que no ens assegura que el model funcioni igual de bé pel Data Set de test.

4.6.1 Accuracy Score

Anomenem *accuracy score* a la puntuació d'encert que té un model a l'hora de predir les mostres de test. Si el nostre model en tot el conjunt de categories predites per una mostra coincideix estrictament amb el conjunt real de categories anotades, la precisió del subconjunt és equivalent a 1. En cas contrari, si les categories predites no s'ajusten en absolut amb les categories reals, el valor és 0. Per tant, sempre ens mourem en valors d'entre u i zero (Pedregosa et al., 2011). Calculem l'*accuracy* mitjançant la següent fórmula:

$$\texttt{accuracy}(y,\hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

Aquesta puntuació ens serveix per tenir una guia per avaluar com està funcionant el nostre model, i en competicions i tasques, veure quin dels models presentats s'adequa millor a l'anàlisi del corpus en qüestió, com per exemple la PAN al CLEF del 2013. El càlcul no discrimina en quines situacions concretes falla el model, simplement fa un càlcul general. Per tal d'analitzar correctament i treure conclusions valoratives, farem servir el que s'anomena Matriu de Confusió, que explicarem a la Secció 4.6.2.

4.6.2 *Matriu de confusió*

La matriu de confusió és una taula en què les files fan referència a les prediccions que es fan de cada categoria i les columnes representen les categories reals. A la Taula 7 hem reproduït un esquema de matriu que representa el funcionament de la Matriu de Confusió que treu el nostre classificador.

Com podem observar, a la Taula 7 tenim unes caselles amb uns valors. A les caselles superiors s'indica les categories reals (que el classificador no coneix) i a les caselles de l'esquerra es representen les categories predites (aqueslles que el nostre classificador ha hagut de predir). En base a aquestes caselles, omplim la resta de caselles. Per tant, la casella amb el valor 1

correspondria a les prediccions que el nostre model ha fet de '*female*' que realment corresponien a la categoria '*female*'. La número 2 són aquelles prediccions que hem fet com a '*female*', però realment eren '*male*'. La casella 3 són aquelles prediccions que el model ha fet com '*male*', però són '*female*' i, per últim, la casella 4 correspon al nombre de prediccions que el model ha efectuat com '*male*' i realment corresponien a la categoria '*male*'.

		REAL	
		female	male
PREDIT	female	1	2
	male	3	4

Taula 7: Esquema d'una Matriu de Confusió

4.6.3 Rànquing PAN al CLEF del 2013

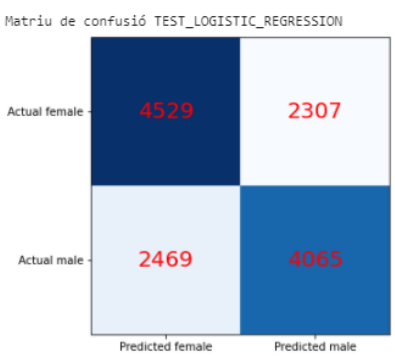
En tota competició s'acostuma a treure un article sobre les tasques, i habitualment, s'acompanya d'una documentació de les tècniques emprades i els resultats que s'han tret en cada tasca, com hem referenciat a la Secció 3, sobre la PAN al CLEF. En la competició que es va dur a terme en el 2013 i de la qual hem emprat el corpus CLEF2013, es van publicar els resultats amb una taula en forma de rànquing (Rangel et al., 2013) on es disposaven de primer a últim els equips participants amb les seves puntuacions, en cert punt de la taula, en concret en el punt on l'*accuracy score* era de 0.5, es va traçar una línia on a partir d'allà cap abaix, es considerava una classificació aleatoria. Hem volgut aprofitar aquesta taula, i hem replicat la taula de la tasca de la classificació de gènere per posar la nostra *accuracy score* per veure en quin lloc hauríem quedat d'haver competit en el seu dia. Els resultats que hem tret no queden molt lluny dels que es van treure, de fet, hauríem quedat en tercer lloc després de l'equip Pastor L. (López-Monroy et al., 2013) i l'equip Santosh (Santosh et al., 2013), com podem veure a la Taula 8.

Team	Gender
Pastor L.	0.6558
Santosh	0.6430
David Cabestany	0.6427
Cruz	0.6219
Flekova	0.5966
Ladra	0.5727
De-Arteaga	0.5429
Kern	0.5375

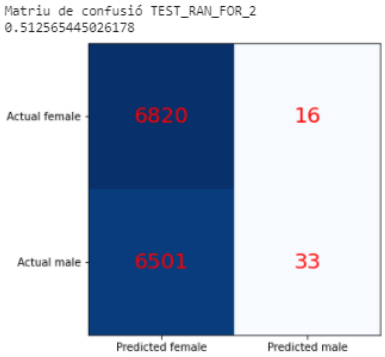
Taula 8: Rànquing de resultats en la tasca de gènere

5 ANÀLISI DELS RESULTATS

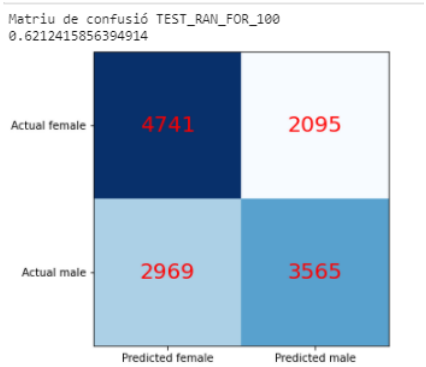
Per evaluar els dos models utilitzats hem generat 4 matrius de confusió (Figura 4). La Figura 4a correspon als resultats que hem extret en el nostre model de Regressió Logística, la Figura 4b representa els resultats que hem obtingut amb el Random Forest de 2 arbres; La Figura 4c són els resultats del Random Forest de 100 arbres i la Figura 4d el de 800 arbres.



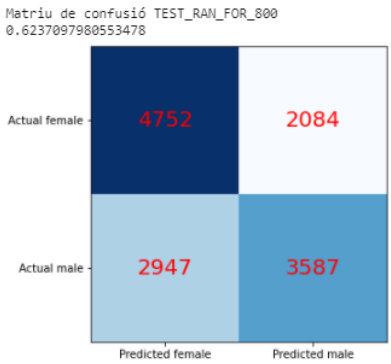
(a) Matriu de confusió Logistic Regression



(b) Matriu de confusió Random Forest 2



(c) Matriu de confusió Random Forest 100



(d) Matriu de confusió Random Forest 800

Figura 4: Matrius de confusió dels models entrenats

En la matriu de confusió que tenim a la Figura 4a veiem que la Regressió Logística ha predit correctament 4529 '*female*' i 4065 '*male*'. En principi no hi ha tanta confusió del model, és a dir, les prediccions més voluminoses corresponen a la categoria correcta, tot i que encara hi ha textos que classifica malament.

En la matriu 4b, 4c i 4d, observem els resultats obtinguts en funció de la quantitat d'arbres utilitzats, tal com hem vist a la Secció 4.4.2. A la Figura 4b veiem que la predicció del model Random Forest amb dos arbres de decisió ha classificat quasi tots els textos dins la categoria '*female*' tret de 16 '*female*' que ha classificat com a '*male*' i 33 '*male*' que ha classificat

correctament. Això fa que s'obtingui una *accuracy score* de 0.51. Una puntuació tan propera a 0.50 no és representativa, ja que es considera una predicció arbitrària. Quan ens trobem en aquesta situació diem que el model s'ha confós. El model de Random Forest amb 100 (Figura 4c) i 800 arbres (Figura 4d) observem uns resultats més adients al que esperàvem. En aquestes Matrius de Confusió observem que l'*accuracy score* puja lleugerament a la de 800 arbres respecte a la de 100 arbres, però en comparació a l'increment de temps que hi ha respecte a la predicció del Random Forest de 100 arbres, aquesta diferència no és rellevant.

El nostre classificador amb Regressió Logística després d'haver estat entrenat amb el corpus CLEF2013 ha aconseguit una *accuracy score* de 0,6427. Aquest valor, tot i ser el tercer millor valor en la classificació, és un valor baix, el que vol dir que hem de continuar posant a prova el classificador, ja que només es troba 14 punts per sobre d'una classificació arbitrària del gènere dels autors. El millor resultat l'hem obtingut amb la Regressió Logística. Sembla que la Regressió Logística funciona bastant bé amb categories binominals, almenys, lleugerament millor que els Random Forests, en què fins i tot la prova amb 800 arbres ha donat una puntuació inferior.

6 CONCLUSIONS I LÍNIES FUTURES

Tenint en compte els resultats que hem obtingut del nostre classificador de gènere amb els dos models que hem implementat, podem deduir que no és aconsellable predir els perfils d'autoria basant-nos tan sols en la freqüència d'aparició de paraules. Si volem que el classificador obtingui uns millors resultats, haurem de fer una cerca dels textos en què ha fallat i a què es deuen aquestes errades, és a dir, una anàlisi d'errors més detallada.

Tot i que creiem que s'haurien de tenir en compte més factors lingüístics, com per exemple, la funció gramatical de cada paraula, la implicació cultural de la paraula, com per exemple, l'ús d'estereotips en el discurs i els argots, paraules especialitzades per nombrar certs tipus d'objectes, o freqüències de paraules amb funcions específiques com els pronoms i veure si la manera d'emprar-los és distintiu en homes o dones. Recordem que si es volguessin tenir en compte els pronoms, primer els hauríem d'excloure del diccionari d'*stop words*.

Comunament l'ús de les majúscules en un comentari en xarxes socials són atribuïdes a alçar la veu. Pensem que d'alguna manera podrien interpretar-se com una forma inconscient d'imposar-se davant l'interlocutor, per tant, és possible que tenir-les en compte en experiments futurs pot ser quelcom interessant.

De la mateixa manera podríem considerar l'ús de col·loquialismes, insults, eufemismes, i altres adjectius qualificatius en discursos i veure si hi ha una relació directa en el gènere de l'autor del text. Una altra qüestió a tenir en compte podria ser la freqüència de les paraules segons la seva funció gramatical fent, prèviament, un PoS tagging dels textos, i fins i tot, podríem tenir en compte la llargada de les paraules que s'empren, per veure si els homes o les dones

mostren una tendència a utilitzar paraules més llargues o més curtes.

Voldríem fer notar que en el corpus CLEF2013, probablement degut al caràcter informal del textos, hem trobat un ús molt estès d'emoticones que en el nostre classificador no hem pogut tenir en compte per falta de temps. Creiem que seria interessant fer-ho d'alguna manera en futures rèpliques, encara que sabem que hi van haver participants en la PAN at CLEF del 2013 que van tenir-les en compte en qüestions de freqüència, i les van descartar posteriorment. Nosaltres podríem tenir en compte els sentiments associats a aquestes emoticones i fer una anàlisi en base a aquest sentiment. Ens hagués agradat tenir en compte les faltes d'ortografia i veure si algun perfil d'autor té més tendència a apropar-se a l'escriptura estàndard que l'altre.

Pel que fa a línies futures del classificador, ens agradaria fer una implementació de diverses tècniques de PLN per veure com es comporten en la classificació de textos segons el gènere de l'autor. Entre aquestes tècniques voldríem posar a prova el classificador amb altres tècniques que tendeixen a utilitzar-se avui en dia (Secció 3). És a dir, ens interessaria provar com es comporten les tècniques *context-dependent* com els *transformers* tipus BERT, així com tècniques de Deep Learning com les xarxes neuronals recurrents (RNN) i les convolucionals (CNN), i altres tècniques d'Aprenentatge Automàtic No Supervisat.

Per últim, volem fer notar que aquest tipus de classificadors i de les diferents tècniques de creació automàtica de perfils encara no s'han dissenyat pel català. Pensem que seria un bon punt a tenir en compte per línies futures de recerca.

La creació de corpus etiquetats amb dades demogràfiques en temàtiques discursives en diferents àmbits seria quelcom útil per poder crear diversos models de detecció d'autoria i de *profiling* en català, per tal d'implementar-los en disciplines com les que hem enumerat en la Secció 1 del nostre treball, per exemple, els estudis de màrqueting, la lingüística forense, el frau i la ciberseguretat, la moderació de continguts, així com la identificació de perfils i grups protegits en xarxes socials de manera automàtica.

Referències

Alemán, Y., Loya, N., Vilariño, D., & Pinto, D. (2013). Two Methodologies Applied to the Author Profiling Task. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, *52*(2), 119–123.

Bedós, C. G. (2020). ¿ escriben igual hombres y mujeres? un estudio de los generolectos en la comunicación mediada por ordenador. *Triangle*(15), 1–49.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *CoRR*, *abs/2005.14165*. Retrieved from https://arxiv.org/abs/2005.14165

Caurcel Díaz, A., & Gómez Hidalgo, J. (2013). Experiments with SMS Translation and Stochastic Gradient Descent in Spanish Text Author Profiling—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Chambers, J. K., & Schilling, N. (2018). *The handbook of language variation and change*. John Wiley & Sons.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, *abs/1406.1078*. Retrieved from http://arxiv.org/abs/1406.1078

Córdoba, M. E. (2020). El constructivismo sociocultural lingüístico como teoría pedagógica de soporte para los estudios generales. *Revista Nuevo Humanismo*, *8*(1).

Cruz, F., Haro R., R., & Ortega, F. (2013). ITALICA at PAN 2013: An Ensemble Learning Approach to Author Profiling—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

D. Weren, E., Moreira, V., & de Oliveira, J. (2013). Using Simple Content Features for the Author Profiling Task—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

De-Arteaga, M., Jimenez, S., Dueñas, G., Mancera, S., & Baquero, J. (2013). Author Profiling Using Corpus Statistics, Lexicons and Stylistic Features—Notebook for PAN at CLEF-2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. Retrieved from

http://arxiv.org/abs/1810.04805

Farrell, A., McDevitt, J., Bailey, L., Andresen, C., & Pierce, E. (2004). *Massachusetts racial and gender profiling study: final report.* (Tech. Rep.).

Flekova, L., & Gurevych, I. (2013). Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Gillam, L. (2013). Readability for author profiling?—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020, September). Array programming with NumPy. *Nature*, *585*(7825), 357–362. Retrieved from https://doi.org/10.1038/s41586-020-2649-2 doi: 10.1038/s41586-020-2649-2

Hernández, D.-I., Guzmán-Cabrera, R., Reyes, A., & Rocha, M.-A. (2013). Semantic-based Features for Author Profiling Identification: First insights—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Holmes, J. (1992). Women's talk in public contexts. *Discourse & Society*, *3*(2), 131–150.

Iyyer, M., Manjunatha, V., Boyd-Graber, J., & Daumé III, H. (2015). Deep unordered composition rivals syntactic methods for text classification. In *Association for computational linguistics.*

Jones, D. (1980). Gossip: Notes on women's oral culture. *Women's studies international quarterly*, *3*(2-3), 193–198.

Lakoff, R. (1973). Language and woman's place. *Language in society*, *2*(1), 45–79.

Lim, W.-Y., Goh, J., & Thing, V. (2013). Content-centric age and gender profiling—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

López-Monroy, A., y Gómez, M. M., Escalante, H., Villaseñor-Pineda, L., & Villatoro-Tello, E. (2013). INAOE's participation at PAN'13: Author Profiling task—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Lozano Domingo, I. (1995). Lenguaje femenino, lenguaje masculino. *Madrid: Minerva Ediciones*.

Mechti, S., Jaoua, M., Belguith, L., , & Faiz, R. (2013). Author Profiling Using Style-based Features—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.),

CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain. CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Meina, M., Brodzińska, K., Celmer, B., Czoków, M., Patera, M., Pezacki, J., & Wilk, M. (2013). Ensemble-based Classification for Author Profiling Using Various Features—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Moreau, E., & Vogel, C. (2013). Style-based Distance Features for Author Profiling—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Mouton, P. G. (2003). *Así hablan las mujeres: curiosidades y tópicos del uso femenino del lenguaje*. LA Esfera De Los Libros SL.

Patra, B., Banerjee, S., Das, D., Saikh, T., & Bandyopadhyay, S. (2013). Automatic Author Profiling Based on Linguistic and Stylistic Features—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Pavan, A., Mogadala, A., & Varma, V. (2013). Author Profiling Using LDA and Maximum Entropy—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54*(1), 547–577.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, June). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N18-1202 doi: 10.18653/v1/N18-1202

Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., & Basile, V. (2019). Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *6th italian conference on computational linguistics, clic-it 2019* (Vol. 2481, pp. 1–6).

Python package index - pypi. (2003). Python Software Foundation. Retrieved 2021-03-28, from https://pypi.org/

Ramos, J., et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, pp. 29–48).

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., & Inches, G. (2013). Overview of the author profiling task at pan 2013. In *Clef conference on multilingual and multimodal information*

access evaluation (pp. 352–365).

Rangel, F., Rosso, P., Potthast, M., & Stein, B. (2017). Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working notes papers of the CLEF*, 1613–0073.

Rangel, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015). Overview of the 3rd author profiling task at pan 2015. In *Clef* (p. 2015).

Rangel, F., Rosso, P., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., ... others (2014). Overview of the 2nd author profiling task at pan 2014. In *Ceur workshop proceedings* (Vol. 1180, pp. 898–927).

Santosh, K., Bansal, R., Shekhar, M., & Varma, V. (2013). Author Profiling: Predicting Age and Gender from Blogs—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Sapkota, U., Solorio, T., y Gómez, M. M., & de-la Rosa, G. R. (2013). Author Profiling for English and Spanish Text—Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.* CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1179

Strand, E. A. (2000). *Gender stereotype effects in speech processing* (Unpublished doctoral dissertation). The Ohio State University.

Vygotski, L. S., Kozulin, A., & Abadía, P. T. (1995). *Pensamiento y lenguaje*. Paidós Barcelona.

Wes McKinney. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (p. 56 - 61). doi: 10.25080/Majora-92bf1922-00a

Wikipedia. (2013). *Tf-idf.* Retrieved from https://ca.wikipedia.org/wiki/Tf-idf

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.emnlp-demos.6 doi: 10.18653/v1/2020.emnlp-demos.6

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1-4), 43–52.