DETECCIÓ DEL LLENGUATGE SEXISTA AMB MODELS DE LLENGUATGE AVANÇATS

TREBALL FI DE GRAU DE

Mireia Carbó Feliu

Director: Horacio Saggion

Grau en Enginyeria matemàtica en ciència de dades

Curs 2023-2024

Universitat Pompeu Fabra Barcelona

Escola d'Enginyeria

Als que sempre em dediqueu un somriure

Agraïments

M'agradaria expressar el meu agraïment a totes les persones que m'han donat la mà en aquest any tan especial, ple de canvis i reptes personals.

En primer lloc, a vosaltres papis, pel vostre suport incondicional a cada decisió. Per educar-me en els valors de l'esforç i ensenyar-me a lluitar pels meus somnis. Marc i Georgina, gràcies per ser cada dia, per no deixar-me caure i espentar-me a continuar sempre amb un gran somriure. Sou llum.

A tots que els que heu format part de la meva etapa universitària, gràcies per compartir tants bons moments. Sense el vostre suport i companyia no hauria estat igual. Us trobo a faltar.

A les de tota la vida, Maries i Vero, gràcies per seguir sempre al meu costat.

Javi, gràcies per la teva ajuda i paciència en tot el procés.

També vull agrair a la Pompeu i al meu tutor, Horacio, per oferir-me l'oportunitat de realitzar aquest treball i proporcionar-me un entorn d'aprenentatge enriquidor.

A tots vosaltres, gràcies de tot cor. Aquest treball no hauria estat possible sense cadascun de vosaltres.

Resum

L'ús excessiu i inadequat de les xarxes socials pot propiciar a la propagació d'actituds sexistes i discriminatòries, afectant diversos col·lectius vulnerables. És important reconèixer que aquestes plataformes contenen contingut sexista, podent comportar a situacions socials greus com la violència de gènere, discriminació laboral, objectificació i assetjament.

L'objectiu principal d'aquest treball és investigar i desenvolupar programes capaços de detectar textos sexistes en tres idiomes: anglès, castellà i turc, categoritzant-los segons la seva tipologia: ideologia i desigualtat, estereotips i domini, misogínia i violència no-sexual, objectificació i violència sexual.

Per dur a terme aquest projecte, s'utilitzen diversos models de classificació simples i models de processament de llenguatge natural a gran escala, els quals detecten característiques clau i analitzen el context semàntic per classificar els textos sexistes. S'utilitzen tant models d'aprenentatge automàtic, com models monolingües i multilingües prèviament entrenats, amb l'objectiu de fer un anàlisis exhaustiu i comparatiu en diferents llengües.

Resumen

El uso excesivo e inadecuado de las redes sociales puede propiciar la propagación de actitudes sexistas y discriminatorias, afectando a varios colectivos vulnerables. Es importante reconocer que estas plataformas contienen contenido sexista, pudiendo comportar situaciones sociales graves como la violencia de género, discriminación laboral, objetificación y acoso.

El objetivo principal de este trabajo es investigar y desarrollar programas capaces de detectar textos sexistas en tres idiomas: inglés, español y turco, categorizándolos según su tipología: ideología y desigualdad, estereotipos y dominio, misoginia y violencia no-sexual, objetificación y violencia sexual.

Para llevar a cabo este proyecto, se utilizan diferentes modelos de clasificación simples y modelos de procesamiento de lenguaje natural a gran escala, que detectan características clave y analizan el contexto semántico para clasificar los textos sexistas. Se utilizan tanto modelos de aprendizaje automático, como modelos monolingües y multilingües previamente entrenados, con el objetivo de realizar un análisis exhaustivo y comparativo en diferentes lenguas.

Abstract

Excessive and inappropriate use of social media can lead to the spread of sexist and discriminatory attitudes, affecting various vulnerable groups. It is important to recognise that these platforms contain sexist content, which can lead to serious social situations such as gender violence, labour discrimination, objectification and harassment.

The main objective of this work is to research and develop software able to detect sexist texts in three languages: English, Spanish and Turkish, categorising them according to their typology: ideology and inequality, stereotypes and dominance, misogyny and non-sexual violence, objectification and sexual violence.

To carry out this project, several simple classification models and large-scale natural language processing models are used, which detect key features and analyze the semantic context to classify sexist texts. Both machine learning models and pre-trained monolingual and multilingual models are used in order to perform a comprehensive and comparative analysis in different languages.

Índex del treball

Introducció **1**
1.1 Contextualització 1
1.2 Objectius 1
1.3 Estructura del treball 2
Estat de l'art **3**
2.1 Sexisme a les xarxes socials 3
2.2 Projecte EXIST 3
2.3 Moviments socials 4
2.4 Estudis previs 5
Marc teòric **9**
3.1 Definició del sexisme 9
3.2 Tipologies de sexisme 9
3.3 Impacte social 10
Metodologia **13**
4.1 Descripció de les dades 13
4.2 Processament de les dades 16
4.3 Llibreries 18
4.4 Eines i tècniques PLN 18
4.5 Models de classificació 19
4.6 Models de PLN 21
4.7 Entrenament dels models aplicats 22
4.8 Avaluació dels models 24
Anàlisi i resultats **27**
5.1 Interpretació dels resultats 27
5.2 Limitacions de l'estudi 32
Interfície web interactiva **34**
6.1 Arquitectura 34
6.2 Funcionament de la interfície 34
Conclusions **35**
Projectes futurs **36**
Referències bibliogràfiques **37**
Apèndix 1: Resultats models de classificació **43**
Apèndix 2: Exemple interfície web **48**
Apèndix 3: Connexió al HPC **50**
Apèndix 4: Abreviatures **52**

Llista de figures

Figura 1: Esquema pipelines 23

Figura 2: Matriu de confusió 24

Figura 3: Resultats de la mètrica loss per a la tasca 1 en el model BERT emprant versions específiques per idioma 29

Figura 4: Resultats de la mètrica loss per a la tasca 2 en el model BERT emprant versions específiques per idioma 29

Figura 5: Resultats de la mètrica loss per a la tasca 1 en el model BERT emprant la versió multilingüe 29

Figura 6: Resultats de la mètrica loss per a la tasca 2 en el model BERT emprant la versió multilingüe 29

Figura 7: Resultats de la mètrica loss en el model BERT emprant la versió multilingüe i el conjunt de dades de tots els idiomes 30

Figura 8: Resultats de la métrica accuracy per al model BERT emprant versions específiques per idioma 30

Figura 9: Resultats de la mètrica accuracy per al model BERT emprant la versió multilingüe 31

Figura 10: Resultats de la mètrica loss per a la tasca 1 en el model RoBERTa emprant versions específiques per idioma 31

Figura 11: Resultats de la mètrica loss per a la tasca 2 en el model RoBERTa emprant versions específiques per idioma 31

Figura 12: Resultats de la mètrica accuracy per al model RoBERTa emprant versions específiques per idioma 32

Figura 13: Resultats de la métrica accuracy per al model XLM RoBERTa 32

Figura 14: Pàgina d'inici 48

Figura 15: Pàgina de classificació 48

Figura 16: Pàgina de classificació amb resultats del comentari 49

Llista de taules

Taula 1: Exemple del conjunt de dades exist 2021 14

Taula 2: Exemple del conjunt de dades exist 2024 15

Taula 3: Exemple del conjunt de dades en turc 15

Taula 4: Instàncies per categoria del conjunt de dades d'entrenament 16

Taula 5: Estructura del conjunt de dades final 16

Taula 6: Resultats del model Naive Bayes per a la tasca 1 43

Taula 7: Resultats del model Naive Bayes per a la tasca 2 43

Taula 8: Resultats del model KNN per a la tasca 1 44

Taula 9: Resultats del model KNN per a la tasca 2 44

Taula 10: Resultats del model Random forest per a la tasca 1 45

Taula 11: Resultats del model Random Forest per a la tasca 2 45

Taula 12: Resultats del model Gadient boosting per a la tasca 1 46

Taula 13: Resultats del model Gadient boosting per a la tasca 2 46

Taula 14: Resultats del model Regressió logística per a la tasca 1 47

Taula 15: Resultats del model Regressió logística per a la tasca 2 47

Taula 16: Abreviatures del treball 53

Capítol 1

Introducció

Avui en dia les xarxes socials juguen un paper central en la vida quotidiana de les persones, transformant la manera de comunicar, interactuar i consumir informació. Aquestes plataformes permeten tenir una comunicació instantània i global, trencant barreres culturals i geogràfiques.

El sexisme dintre de les xarxes socials es manifesta de diverses maneres. Des de comentaris despectius i llenguatge ofensiu fins a la difusió d'estereotips de gènere[1] i la violència de gènere digital, les xarxes socials han esdevingut un espai on les dinàmiques de poder i desigualtat de gènere es reprodueixen i, de vegades, s'intensifiquen.

1.1 Contextualització

L'estudi del sexisme a les xarxes socials és crucial. La influència de les xarxes socials en la percepció i la perpetuació de rols[2] i estereotips de gènere és significativa. Estudis han demostrat que el contingut sexista i la misogínia[3] en línia poden tenir efectes perjudicials en l'autoestima, la salut mental i el benestar de les persones.

L'anonimat i l'accessibilitat que ofereixen poden contribuir a la proliferació de comportaments sexistes. En aquest context és crucial desenvolupar eines i models intel·ligents que permeten identificar, classificar i mitigar aquests comportaments a les xarxes socials.

1.2 Objectius

L'objectiu d'aquest treball final de grau consisteix a investigar, desenvolupar, analitzar i avaluar el rendiment de diferents models de llenguatge natural per a la detecció i classificació de textos sexistes en tres idiomes: anglès, castellà i turc.

Per assolir aquest objectiu general, en primer lloc, hem obtenir una font de dades fiables en els diferents idiomes. Aquestes dades han d'estar ben etiquetades i categoritzades tal com veurem al llarg del treball. La qualitat de les dades és fonamental per a poder dur a terme un entrenament efectiu dels models.

Abans de començar, és important fer un recerca i aprofundir en el coneixement dels models de llenguatge natural que siguin capaços d'analitzar aquestes dades. Avaluarem diverses arquitectures i enfocaments, des de models d'aprenentatge automàtic tradicionals, a models monolingües i multilingües prèviament entrenats.

Un cop tinguem les dades, entrenarem els models, ajustant els paràmetres per optimitzar el rendiment. I posteriorment es realitzarà una avaluació exhaustiva dels resultats amb diferents mètriques. Aquesta anàlisi permetrà identificar les fortaleses i debilitats de cada enfocament, proporcionant una base sòlida per a futures investigacions i desenvolupaments en la detecció de sexisme a les xarxes socials.

[1] Un estereotip de gènere és una creença o idea generalitzada sobre les característiques o rols que les dones i els homes haurien de tenir o exercir [1].
[2] La perpetuació dels rols de gènere és el procés continu pel qual la societat manté i reforça les expectatives sobre els diferents rols que haurien de tenir les dones i els homes[2].
[3] La misogínia és l'odi arrelat a les dones[3].

1.3 Estructura del treball

Per facilitar la comprensió i seguiment del treball, a continuació, es descriu breument el contingut de cada secció:

- **Introducció**: Aquest apartat ofereix una visió del context en el qual es desenvolupa el treball, així com els objectius i l'estructura que segueix.
- **Estat de l'art**: Aquest apartat presenta una anàlisi sobre el sexisme a les xarxes socials, examinant moviments socials i estudis previs.
- **Marc teòric**: Es proporciona una definició clara i precisa del sexisme, i les tipologies de categorització utilitzades al llarg del treball. A més, es du a terme una anàlisi de l'impacte del sexisme a la societat, abordant-ne les conseqüències a nivell individual i col·lectiu.
- **Metodologia**: Aquest apartat ens detalla la naturalesa i l'origen de les dades utilitzades en la investigació. Així com el processament de dades que s'ha realitzat.
  També trobem els models de classificació i de Processament del Llenguatge Natural (PLN) utilitzats a l'estudi. Finalment, es detalla el procés d'entrenament dels models, incloent-hi les mètriques d'avaluació utilitzades per avaluar el rendiment dels models.
- **Anàlisi i resultats**: En aquest apartat trobem la interpretació i discussió dels resultats. També les limitacions que s'han trobat en la realització del treball.
- **Web interactiva**: Per a poder visualitzar la classificació d'un text introduït per un usuari.
- **Conclusions**: Resumeix les principals troballes de la investigació, destacant-ne la rellevància i les aportacions al coneixement existent.
- **Projectes futurs**: En aquest apartat se suggereixen línies de recerca futures, basades en els resultats i les limitacions de l'estudi actual.
- **Referències bibliogràfiques**: Trobem una llista completa de les fonts i referències consultades al llarg de la investigació.
- **Annexos**: En aquest apartat trobem informació addicional del treball.

Capítol 2

Estat de l'art

En aquesta secció es proporciona la informació més rellevant i recent sobre el sexisme a les xarxes socials. Això ens permet situar la investigació en el context adequat.

La secció està estructurada en quatre parts. La primera examina com el sexisme es manifesta i es perpetua a les xarxes. Seguida del Projecte EXIST, una iniciativa que utilitza tecnologies avançades com la intel·ligència artificial i el processament del llenguatge natural per detectar i analitzar el sexisme a les xarxes socials. També trobem moviments socials que mobilitzen i visibilitzen el sexisme, així com el seu impacte a la consciència pública i els canvis socials que han impulsat. Finalment, es fa una revisió dels estudis existents que aborden aquest tema.

2.1 Sexisme a les xarxes socials

Les xarxes socials s'han convertit en una eina essencial en la nostra vida diària. Tanmateix, també reflecteixen les desigualtats i dinàmiques de poder existent en la societat, on s'inclou el sexisme.

Per una banda, les xarxes socials permeten la ràpida difusió de contingut, afavorint la difusió de contingut sobre la igualtat[4] i els drets humans[5], conscienciant sobre les diverses formes de discriminació o fomentant la participació de moviments feministes[6]. No obstant això, també es converteixen en escenaris d'assetjament i atacs sistemàtics. A més, aquestes plataformes presenten un alt contingut de comentaris sexistes i ofensius que perpetuen estereotips i arriben a crear un ambient hostil.

L'anonimat a les xarxes socials fomenta comportaments abusius, ja que els usuaris poden difondre contingut perjudicial sense por de repercussions. La manca de regulació i l'aplicació inconsistent de polítiques dificulten la prevenció i el maneig d'aquests problemes.

En resum, encara que les xarxes socials poden ser eines poderoses per a la visibilització, la protesta i la denúncia, també presenten desafiaments significatius en la lluita contra el sexisme. És essencial desenvolupar estratègies i polítiques per crear un entorn més segur i equitatiu per a tots els usuaris.

2.2 Projecte EXIST

El projecte EXIST (*sEXism Identification in Social networks*) se centra en la detecció del sexisme a les xarxes socials. Aquesta iniciativa aborda el problema del sexisme utilitzant tecnologies avançades com la intel·ligència artificial i el processament del llenguatge natural. El seu objectiu principal és detectar, analitzar i comprendre les diverses maneres en què es manifesta el sexisme en un sentit ampli, des de la misogínia explícita fins a altres expressions subtils que impliquen comportaments sexistes implícits.

El projecte ha tingut diverses edicions. La primera edició d'EXIST es va realitzar com a tasca compartida amb CLEF a l'any 2021. El seu objectiu principal va ser detectar el sexisme a les xarxes socials, definint dues tasques principals. En primer lloc, una tasca de detecció, el propòsit de la qual era identificar automàticament instàncies de sexisme en textos de xarxes socials, classificant els textos com a sexistes

[4] La igualtat és el principi que totes les persones tenen dret a les mateixes oportunitats i tracte, independentment del seu gènere, raça, ètnia, religió, orientació sexual o qualsevol altra característica [4].
[5] Els drets humans són els drets i llibertats bàsics que pertanyen a totes les persones al món, des del naixement fins a la mort [5].
[6] El moviment feminista és un conjunt divers de moviments socials, polítics, econòmics i culturals que tenen com a objectiu definir i establir la igualtat política, econòmica, personal i social dels sexes [6].

o no sexistes. En segon lloc, una tasca de classificació, on es categoritzaven les instàncies sexistes en les categories següents: ideologia i desigualtat, estereotips i domini, misogínia i violència no sexual, objectificació i violència sexual. Aquesta classificació permet una anàlisi més detallada de les diverses formes en què el sexisme es manifesta i es percep en línia [7].

La segona edició va continuar amb el mateix enfocament i tasques, centrant-se a categoritzar el sexisme segons la intenció de l'autor. A la tercera edició, que es va dur a terme com un laboratori a CLEF 2023 a Tessalònica, Grècia, es van mantenir la classificació binària i la categorització de la intenció de l'autor, afegint-se una nova tasca: la identificació del sexisme enfocada a la intenció de la font original del comentari [8].

L'edició d'EXIST 2024 comparteix el propòsit de detectar i classificar el sexisme a les xarxes socials com les tres edicions anteriors, però en aquest cas la base de dades inclou *tweets* i *memes*. Per tant, s'hi incorporen noves tasques centrades en imatges, especialment *memes*. Els *memes* són imatges, generalment de naturalesa humorística, que es propaguen ràpidament a les xarxes socials i Internet. L'objectiu és abastar un espectre més ampli de manifestacions sexistes a les xarxes socials, especialment aquelles disfressades d'humor [9].

Per tant, el projecte EXIST ha demostrat ser una iniciativa crucial en la identificació i comprensió del sexisme a les xarxes socials al llarg de les seves diferents edicions. En emprar tecnologies avançades i ampliar-ne contínuament l'abast i els mètodes d'anàlisi, EXIST aborda de manera efectiva les diverses manifestacions del sexisme, adaptant-se a noves formes de comunicació digital com els *memes*. Amb cada edició, el projecte no només millora les capacitats de detecció i classificació, sinó que també contribueix significativament a la conscienciació i entesa d'aquest problema social persistent [10].

2.3 Moviments socials

Els moviments socials en les xarxes socials tenen un paper de vital importància i un gran impacte en la societat.

Les xarxes socials són un altaveu per a poder compartir històries i experiències, poden visibilitzar la violència de gènere i generar empatia i sororitat per les víctimes. Amb l'ajuda dels *hashtags* i campanyes virals s'ha augmentat la consciència pública sobre temes sexistes, com l'assetjament sexual[7], discriminació de gènere i violència.

La visibilitat de les xarxes permet generar pressió per a les institucions i es prenguin mesures cap a la igualtat en lleis o pràctiques institucionals.

A continuació, veiem diversos moviments socials que s'han fet els últims anys.

2.3.1 Marxa Mundial de les Dones

És un moviment social que realitza accions feministes, reunint a diferents grups de dones i organitzacions que lluiten per eliminar la pobresa i violència cap a les dones. Té els seus inicis en 1995, però no és fins a l'any 2000 que realitzen la seva primera acció.

La Marxa Mundial de les Dones utilitza les xarxes socials com un poderós altaveu per a les seves reivindicacions i per connectar dones de tot el món en la seva lluita per un món més equitatiu i segur [12].

[7] L'assetjament sexual es defineix com qualsevol classe de conducta de naturalesa sexual no desitjada que pot ser verbal, no verbal o física i que té com a objectiu o efecte atent contra la dignitat d'una persona, en particular quan es crea un entorn intimidant, hostil, degradant, humiliant o ofensiu [11].

2.3.2 #MeToo

El moviment *#MeToo* sorgeix com a protesta contra l'assetjament i la violència sexual. El 2017 es va originar arran dels casos d'assetjament sexual del productor de Hollywood Harvey Weinstein. I des de llavors el *hashtag #MeToo* ha estat utilitzat per compartir i denunciar violència sexual a les xarxes [13][14].

Aquest moviment ha generat una major consciència sobre la importància de creure i recolzar a les víctimes. A més, ha impulsat canvis en la cultura, empreses i legislació en contra de l'assetjament i la impunitat.

A partir, d'aquest moviment va sorgir *Time's Up*, que promou la lluita contra l'assetjament sexual i promou la igualtat de gènere.

2.3.4 *#BalanceTonPorc* (*#OutYourPig*)

El moviment *#BalanceTonPorc* va ser impulsat per la periodista francesa Sandra Muller, i naix arran de *#MeToo*. Però a diferència de *#MeToo* que es va centrar en figures de gent coneguda i famosos, *#BalanceTonPorc* busca donar veu a dones anònimes que han sigut víctimes d'assetjadors desconeguts en el seu dia a dia [13].

#BalanceTonPorc va sensibilitzar sobre la realitat de l'abús masclista[8] i va mostrar que no només passava al món del cinema. On el primer any es van comptabilitzar més de 931.000 *tweets* amb el *hashtag*, utilitzat per gairebé 300.000 usuaris a Twitter [15].

2.3.5 *#SayHerName*

El moviment *#SayHerName* és una iniciativa que busca crear i conscienciar sobre les dones i nenes negres que són víctimes de la violència antinegra[9] i brutalitat policial als Estats Units.

Aquest moviment va ser crear per l'Associació per a l'Avanç de les Dones i Nenes Negres (AAPF) i el Centre d'Estudis de Política Social i Investigació (CISPS) en desembre de 2014 [16].

Mitjançant aquest *hashtag* a les xarxes socials activistes, periodistes i altres usuaris ajuden a visibilitzar i denunciar els casos, a més de difondre el seu missatge per la lluita de la justícia racial i d'igualtat de gènere.

2.4 Estudis previs

Al llarg dels últims anys s'han realitzat nombrosos estudis relacionats amb el sexisme a les xarxes socials. On s'ha posat de manifest la presència de comportaments i discursos sexistes a diferents plataformes digitals com Twitter, Facebook o Instagram. Aquests estudis identifiquen les diferents formes de sexisme, la propagació d'aquest tipus de contingut i l'impacte social que genera.

Realitzant una recerca d'articles que aborden el sexisme en les xarxes socials, trobem una àmplia quantitat i varietat d'investigacions realitzades per diverses institucions. Destaquem que l'Institut d'Estudis de Gènere de la Universitat Carlos III de Madrid (IEG-UC3M) rep durant tot l'any investigadors/es i professors/es que fan investigacions en l'àmbit dels Estudis de Gènere i Feministes.

[8] L'abús mascclista, també conegut com a violència de gènere contra les dones, és un terme que s'utilitza per descriure qualsevol acte de violència per motius de gènere que cau o pugui causar danys físics, sexuals o psicològics a les dones, incloses les amenaces d'aquests actes, coacció o privació arbitrària de llibertat [17].

[9] La violència antinegra, també coneguda com a violència contra els negres, és un terme utilitzat per descriure qualsevol acte de violència motivat per prejudicis racials contra els negres. Pot ocórrer en àmbits públics o privats, i abasta una àmplia gamma de comportaments [18].

A continuació s'esmenten diferents estudis dels últims anys que aborden el sexisme a les xarxes socials utilitzant tècniques d'aprenentatge automàtic o aprenentatge profund.

2.4.1 Estudis en anglès

Pel que fa als estudis realitzats per a la detecció del sexisme en anglès, trobem una quantitat considerable en comparació de les altres llengües utilitzades en aquest treball. Veiem diferents estudis realitzant en diferents anys:

- Zou et al. (2017): Aquest estudi està centrat en la detecció i classificació de *tweets* sexistes en anglès utilitzant un model basat l'algoritme d'aprenentatge automàtic SVM (Màquines de Suport Vectorial). L'objectiu principal era identificar *tweets* que continguessin llenguatge sexista o discriminatori. L'estudi va aconseguir una precisió del 85%, per tant, el model va ser capaç d'identificar correctament el 85% dels *tweets* sexistes al conjunt de dades utilitzat per a l'avaluació [19].

- Davidson et al. (2018): Aquest estudi s'enfoca en la classificació de *tweets* misògins utilitzant un model basat en xarxa neuronal de tipus LSTM (*Long Short-Term Memory*). Les xarxes LSTM són un tipus de xarxa neuronal recurrent que s'utilitza per processar i classificar seqüències de dades, com a textos. L'estudi va aconseguir un F1-score del 90%, el qual indica que el model va ser capaç d'aconseguir un equilibri entre la precisió i el record en la identificació de tweets misògins [20].

- Yıldırım et al. (2019): En aquest article, Yıldırım i els seus companys van proposar identificar comentaris sexistes a Twitter utilitzant tècniques d'aprenentatge per transferència amb xarxes neuronals convolucionals. L'aprenentatge per transferència és una tècnica on s'aprofiten els coneixements apresos per un model en una tasca per millorar el rendiment en una altra tasca relacionada. L'estudi va aconseguir una precisió del 93%, cosa que indica que el model va ser altament efectiu en la identificació de comentaris sexistes a Twitter [21].

2.4.2 Estudis en castellà

En el cas de castellà també trobem diversos estudis realitzats sobre la detecció del sexisme a les xarxes socials:

- Sanches-Piqueras et al. (2019): Aquest estudi està enfocat en la detecció de missatges sexistes a Twitter en castellà. S'utilitza un model de classificació basat a *Naive Bayes*, amb l'objectiu d'identificar automàticament missatges que contenien llenguatge sexista a la plataforma Twitter. L'estudi va aconseguir una precisió del 78% [22].

- Díaz-Galiano et al. (2020): En el cas d'aquest estudi, l'enfocament de la detecció del sexisme no és a les xarxes socials, sinó als comentaris ofensius en fòrums de notícies en castellà. L'algorisme emprat va ser SVM, on es va aconseguir un bon rendiment amb un F1-score del 82% [23].

- Horta et al. (2021): En aquest estudi, s'identifica els *tweets* sexistes en castellà utilitzant un enfocament d'aprenentatge profund basat en xarxes neuronals recurrents (RNN). El model va aconseguir una precisió del 91%, cosa que suggereix que va ser altament efectiu en la identificació de *tweets* sexistes en castellà [24].

2.4.3 Estudis en turc

Pel que fa a la literatura sobre estudis previs del sexisme a les xarxes socials en turc, s'observa una notable falta d'investigacions i conjunts de dades específiques per dur a terme aquests estudis.

- Lutfiye Seda Mut Altin i Horacio Saggion (2024): Aquest article presenta un nou corpus per a la identificació i categorització automatitzada del sexisme a les xarxes socials en turc, que és l'emprat en aquest treball. En primer lloc, es realitza un estudi previ a la detecció de discurs d'odi a les xarxes socials i se centra en la discriminació de gènere, una àrea menys estudiada.

  En aquest article es presenten els resultats d'experiments realitzats per a la identificació de sexisme en turc, utilitzant diferents models de PLN, com SVM, bi-LSTM i BERT, tant en la seva versió multilingüe com en una adaptada específicament per al turc. Els resultats mostren que els models basats en BERT, especialment l'optimitzat per al turc, van obtenir els millors resultats en la classificació de comentaris a les xarxes socials.

  El conjunt de dades s'ha fet públic per impulsar la investigació en aquesta àrea, remarcant que aquest és el primer conjunt de dades turques exhaustivament anotades per a la identificació de sexisme i suggereixen futures millores, com ara el preprocessament específic del turc, l'ampliació de dades amb models de generació de llenguatge i l'entrenament en models acoblats [25].

- Mansur Alp Tocoglu, Okan Öztürkmenoğlu i Adil Alpkocak van presentar un estudi sobre l'anàlisi d'emocions a *tweets* en turc utilitzant models de xarxes neuronals profundes. A la seva investigació, van examinar tres arquitectures d'aprenentatge profund: xarxes neuronals artificials (ANN), xarxes neuronals convolucionals (CNN) i xarxes neuronals recurrents amb memòria a curt i llarg termini (LSTM).

  Per entrenar aquests models, els autors van crear un conjunt de dades de tweets en turc i els van anotar per classificar-los en sis categories emocionals (alegria, tristesa, ira, por, fàstic i sorpresa) utilitzant un enfocament basat en lèxic. Els resultats obtinguts van mostrar que l'ANN va produir els pitjors resultats, mentre que la CNN va resultar ser la més precisa amb una precisió de 0,74.

  Al llarg de la publicació, es destaca la importància de l'anàlisi de dades a les xarxes socials per entendre les opinions i emocions de les persones, ja que les emocions són fonamentals en la comunicació humana.

  Els autors van contribuir proporcionant un nou conjunt de dades, el *Turkish Twitter Emotion Dataset* (TURTED), que conté més de 195.000 tuits anotats automàticament als sis estats emocionals. Aquest conjunt de dades està disponible per a ús acadèmic. I a més, reivindiquen la manca d'estudis al respecte.

  En els experiments realitzats, es van comparar les arquitectures d'aprenentatge profund amb mètodes tradicionals d'aprenentatge automàtic, demostrant que les xarxes neuronals profundes superen els mètodes tradicionals en l'anàlisi d'emocions al turc. La CNN es va mostrar com l'arquitectura més efectiva, seguida per la LSTM [26].

2.4.4 Altres articles

Ara Zozan Miran i Hazha Saeed Yahia de la Lebanese French University, autors de l'article científic titulat "*Hate Speech Detection in Social Media (Twitter) Using Neural Network*". Publicat al *Journal of Mobile Multimedia* al febrer de 2023. En aquest article es presenta una revisió sistemàtica d'estudis acadèmics sobre la detecció de discurs d'odi a Twitter utilitzant models basats en Xarxes Neuronals Convolucionals (CNN).

A l'article mostra que el discurs d'odi a les xarxes socials es manifesta de moltes maneres, incloent-hi l'abús, l'ofensa, el sexisme, el racisme i l'odi per afiliació política, religiosa, nacionalitat, color de pell, discapacitat, gènere, ètnia, orientació sexual i immigració.

Es desenvolupen diferents mètodes per detectar aquest discurs d'odi. Els resultats obtinguts indiquen que els models basats en CNN són els que obtenen un millor rendiment per a la detecció de discurs d'odi. Tot i això, hi ha problemes, com la incapacitat de molts models per detectar automàticament el discurs d'odi, la inadequació dels mètodes per a tots els idiomes (la majoria funciona millor amb l'anglès), la confusió en la classificació del discurs i la falta de consideració del diàleg entre usuaris a les xarxes socials.

Aquest article també ens mostra que el Regne Unit és el país amb més publicacions sobre detecció de discurs d'odi, seguit per l'Índia i la Xina [27].

En conclusió, malgrat els avenços en la detecció de discurs sexista a les xarxes socials, la revisió sistemàtica ajuda a identificar i desenvolupar nous models que abordin aquests desafiaments.

Capítol 3

Marc teòric

El sexisme abasta una àmplia gamma d'actituds, creences i comportaments que perpetuen la discriminació o el menyspreu. En el marc teòric, es proporciona una definició integral del sexisme, explorant-ne les manifestacions i les implicacions en la societat contemporània. A més, s'examinen les diverses tipologies que s'utilitzen en aquest treball, per tal de comprendre la complexitat d'aquest fenomen i el seu impacte en diferents contextos socials.

A més, s'analitza en profunditat l'impacte social del sexisme, destacant-ne les conseqüències en diferents àmbits.

3.1 Definició del sexisme

El sexisme és el prejudici o discriminació que s'exerceix basant-se en el sexe[10], gènere[11] o identitat sexual[12] de les persones. Englobant tot el conjunt de pràctiques discriminatòries i conductes de desigualtat o estigmatització basades en la creença al voltant del sexe i gènere de les persones, així com a la diversitat d'identitats sexuals.

El terme sexisme sorgeix als Estats Units en la dècada dels seixanta. Aquest terme s'utilitzava per grups feministes per demostrar que el gènere constitueix un factor de discriminació, subordinació i desvaloració cap a les dones [31].

Avui en dia el sexisme es manifesta en diverses àrees i accions. Ja sigui d'una manera subtil als comentaris o actituds que qüestionen la capacitat de les persones que la pateixen, fins a arribar a una violència física, verbal o psicològica. Però també es pot veure el sexisme en els mitjans de comunicació, perpetuant estereotips de gènere i imatges irrealistes, o bé en el món laboral, amb actituds sexistes i desigualtats estructurals [32].

3.2 Tipologies de sexisme

El sexisme és un fenomen complex que es manifesta en diverses formes i contextos. En aquest apartat s'esmenten les diferents tipologies de sexisme utilitzades en tot el treball.

3.2.1. Ideologia i desigualtat

El sexisme presenta una estreta relació amb les creences, valors i ideologies d'una cultura o grup social. On moltes vegades es veu reflectida en la desigualtat de gènere. La desigualtat de gènere és un desequilibri que es deriva de la discriminació basada en el sexe, amb la finalitat de disminuir el reconeixement o drets de les dones.

[10] El concepte sexe fa referència a les característiques biològiques, anatòmiques, fisiològiques i cromosòmiques de l'espècie humana [28].
[11] El gènere es refereix a la identitat social i cultural d'una persona com a home, dona, no binària o altre gènere. El gènere es basa en autoidentificació i no està determinat per les característiques biològiques [29].
[12] La identitat de gènere és el sentit intern d'una persona de ser home, dona, no binària o altre gènere. No és determinada per les característiques biològiques, sinó per la pròpia percepció i experiència de gènere de cada individu [30].

En l'àmbit del llenguatge, la desigualtat es fa evident quan el text rebutja la igualtat entre homes i dones o presenta als homes com a víctimes d'opressió de gènere, així com desacreditant la lluita per la igualtat [33].

3.2.2. Estereotips i dominància

Els estereotips són creences generalitzades sobre els rols i les característiques associades als gèneres. En l'àmbit del llenguatge ho veiem sovint en creences falses sobre les dones assignant, doncs, determinats rols, com podria ser cuidadora, familiar, tendra, submisa, etc. Però també creure en què són inadequades per a realitzar certes tasques com conduir, treballar, etc. Uns estereotips que poden limitar les oportunitats i reforçar els rols tradicionals fruits d'un pensament masclista.

El sexisme també es manifesta a través de la dominància i el control, creant una desigualtat estructural [7][34].

3.2.3. Objectificació

L'objectificació és una forma de tractar a les dones com a objectes separant-les de la seva pròpia humanitat i individualitat. On es pressuposa que les dones han de tenir certes qualitats físiques i complir amb els estàndards de bellesa establerts socialment. A l'àmbit del llenguatge, algunes formes d'objectificació inclouen la hipersexualització, que redueix la dona a la seva aparença i la utilitza com a un objecte per a satisfer el plaer sexual (normalment masculí) o sotmetre-la a unes expectatives físiques [7].

3.2.4. Violència sexual

La violència sexual és qualsevol comportament de naturalesa sexual que es du a terme sense el consentiment de l'altra persona. Incloent conductes com l'exhibicionisme, paraules obscenes, tocament o violació.

En l'àmbit del llenguatge ho podem veure reflectit en suggeriments sexuals, sol·licituds de favors sexuals o assetjament [7][35].

3.2.5. Misogínia i violència no sexual

La misogínia és l'odi, menyspreu o aversió cap a les dones, manifestant-se de diverses maneres utilitzant llenguatge ofensiu i menyspreador cap a les dones. La violència no sexual fa referència a qualsevol acte violent que no té un component sexual directe. En l'àmbit del llenguatge s'observa amb forma d'abús verbal, fent sentir a la víctima pressionada, culpable o avergonyida, o expressant odi i violència cap a dita persona [7][36].

3.3 Impacte social

El sexisme té unes arrels històriques profundes, on la societat s'organitzava al voltant d'estructures patriarcals[13] que atorgaven privilegis i poder als homes, mentre les dones estaven sotmeses a uns rols subordinats i limitats. Fent que aquestes creences perpetuen en les generacions, influint en àrees de la vida social, política, econòmica i cultural [31].

[13] Una estructura patriarcal és un sistema social en què els homes tenen el poder i l'autoritat primaris dins d'una família, comunitat o societat. És un sistema que privilegia els homes sobre les dones i els nens, i sovint inclou normes i costums que dicten els rols i les relacions de gènere [37].

Durant molt de temps el sexisme ha estat present a través de normes i lleis que discriminaven a les dones, com la negació al vot i l'educació. Marcant unes arrels profundes i perjudicials que afecten molts individus i col·lectius.

Avui en dia el sexisme contribueix a la presència de desigualtat de gènere present en aspectes quotidians de la vida, com l'accés en l'educació, el treball, la salut i la participació política. En l'àmbit de les relacions, el sexisme alimenta la violència de gènere i discriminació, normalitzant actituds i comportaments abusius.

En l'àmbit estructural, la falta d'igualtat comporta a situacions de desigualtat d'oportunitats i drets que provoquen exclusió i marginalització [38].

3.3.1. Context històric i cultural

El sexisme és un fenomen complex que es manifesta de manera desigual entre els diferents països i cultures. En Europa, països com Bulgària, Hongria, República Txeca, Polònia i Letònia presenten alt índex de sexisme i normes culturals que estableixen rols de gènere rígids. En canvi, els països nòrdics han avançat molt cap a la igualtat de gènere.

Si es du a terme una anàlisi a escala global, els països menys desenvolupats tendeixen a tenir alts índexs de sexisme, on es veu una clara desigualtat entre homes i dones. Ja que solen tenir normes culturals on els rols de gènere són molt rígids i desiguals [39][40].

3.3.2. Factors socioeconòmics

El factor socioeconòmic juga un paper fonamental en el sexisme. L'índex de desenvolupament humà (IDH) mesura el desenvolupament econòmic, educatiu i de salut d'un país, i proporciona informació de l'índex del sexisme. On països amb un baix IDH tendeixen a ser molt sexistes.

El sexisme afecta directament a la distribució dels ingressos i riquesa entre les persones, amb factors com la bretxa salarial, la segregació ocupacional o la manca d'oportunitats. Provocant una dependència financera caps als homes [41].

3.3.3. Polítiques de gènere

Les polítiques de gènere són estratègies implementades per als governs i organitzacions que promouen la igualtat entre homes i dones.

En juliol de 2022 a Espanya es va aprovar la Llei Integral per a la Igualtat de Tracte i la No Discriminació, amb l'objectiu de garantir la igualtat i prevenir la discriminació en l'àmbit laboral, educació, salut i participació política. La llei prohibeix la discriminació per motius de naixement, raça, sexe, religió, opinió o altres circumstàncies personals o socials [42].

3.3.4. Mitjans de comunicació

El sexisme en els mitjans de comunicació perpetuen la discriminació i els rols estereotipats, com pot ser la figura de l'home poderós i la dona submisa, presents a telenovel·les, campanyes de publicitat o inclús a les notícies [43].

Els mitjans tenen una responsabilitat molt gran a l'hora de crear contingut que eliminen els estereotips masclistes i fomentar un llenguatge inclusiu.

3.3.5. Educació

El sexisme en l'educació pren un paper fonamental a l'hora de transmetre i fomentar la igualtat. Però avui en dia, encara veiem moltes pràctiques pedagògiques, organitzacions, utilització del llenguatge, llibres de text i patrons que reforcen els estereotips de gènere [44].

En l'àmbit de l'educació, s'observen certs patrons en quant a l'elecció d'estudis superiors, on les dones tendeixen a àrees relacionades amb les arts i humanitats, mentre que els homes a enginyeria o tecnologia [45].

La falta de representació de dones en la història i en la ciència, fa que augmenti la creença del fet que les dones no han contribuït significativament. Fent creure a moltes nenes que no tenen lloc en aquest món.

3.3.6. Salut i benestar

El sexisme no és únicament una qüestió de justícia social, sinó que també afecta a la salut i el benestar de les persones. És d'important rellevància tenir una educació sexual completa i lliure de prejudicis. I que aborden temes com la diversitat sexual i el consentiment.

L'exposició constant a actituds sexistes pot contribuir a la depressió, ansietat i problemes greus de salut mental. Així com lesions físiques a causa d'accidents o violència.

D'altra banda, el sexisme pot afectar a l'accés a l'atenció mèdica de les dones, ja que les creences de gènere poden causar subestimació de símptomes o inclús negació a l'atenció [46].

Capítol 4

Metodologia

En l'apartat de metodologia es descriuen els diferents procediments i tècniques que s'han emprat en el treball, i les diferents etapes de les dades utilitzades.

4.1 Descripció de les dades

En aquesta secció detallem la naturalesa i característiques del nostre conjunt de dades utilitzat per entrenar els models d'aprenentatge automàtic.

El conjunt de dades utilitzat està compost per tres conjunts de dades diferents, dos d'ells proporcionats pel concurs EXIST i l'altre per Seda Mut (estudiant de doctorat):

- **EXIST 2021:** El conjunt de dades EXIST incorpora qualsevol tipus d'expressió sexista o fenòmens relacionats, incloses les afirmacions descriptives o informades on el missatge sexista és un informe o una descripció d'un comportament sexista. Aquest conjunt de dades recull expressions i termes populars, tant en anglès com en castellà, extrets de Twitter i Gap.

  Aquests termes van ser analitzats i filtrats per dues expertes en qüestions de gènere, Trinidad Donoso i Miriam Comet, per finalment obtenir un conjunt de dades amb més de 200 expressions que es poden utilitzar en contextos masclistes.

  La recopilació de dades es va portar a terme des de l'1 de desembre de 2020 fins al 28 de febrer de 2021.

  Respecte a les etiquetes, cada *tweet* va ser anotat per 5 anotadors de *crowdsourcing*[14] seguint les directrius de les expertes en qüestions de gènere i es va dur a terme una prova d'acord entre els anotadors, on les etiquetes finals corresponen al vot majoritari. [7]

text	language	task1	task2
She calls herself "anti-feminazi" how about shut the fucking up on your vile commentary on an elderly responsible citizen tu sach muuch ghani baawri-bewdi hai bey https://t.co/ZMxTDwsY5D	en	sexist	ideological-inequality
@nytimes I learned needlepoint to pick up chicks when traveling. #pua #mgtow	en	sexist	misogyny-non-sexual-violence
Now, back to these women, the brave and the beautiful, @Clare_Crawley and @tayshia. These bad ass babes, are deserve so much credit for how this season has gone. As a woman, I've learned so much from them and feel more empowered to expect more in future relationships.	en	non-sexist	non-sexist
@CurvyBandida @Xalynne_B Wow, your skirt is very short. What is it's length? 5 inch or more?	en	sexist	objectification

[14] El crowdsourcing és el procés d'obtenir informació, idees o serveis d'un gran grup de persones, normalment a través d'Internet. És com aprofitar la intel·ligència i les habilitats col·lectives d'una multitud.

@Smithcouple971 Hello....m raj....m with good size and excellent stamina ....A passionate pussy licker...Love to lick every holes were womens desire to b licked...DoggyMissionaryWomen on topMy best pose to spank over bed	en	sexist	sexual-violence
I really just want to be rich but not trophy wife rich, rich with my own	en	sexist	stereotyping-dominance

Taula 1: Exemple del conjunt de dades exist 2021

- **EXIST 2024:** EXIST 2024 pretén perfeccionar i ampliar l'abast de la detecció de sexisme incorporant tant text com *memes*, millorant la robustesa del conjunt de dades amb un lèxic ben curat i un enfocament equilibrat de la recollida i anotació de dades. El conjunt de dades abasta diversos temes, incorporant termes amb diferents graus d'ús tant en contextos sexistes com no sexistes, tots centrats en les dones.

  Cada *meme* va ser anotat per 6 anotadors col·lectius seleccionats mitjançant l'aplicació Prolific[15], seguint les directrius desenvolupades per experts en gènere. Per tal de mitigar el biaix de l'etiqueta, en aquest conjunt, es consideren dos paràmetres socials i demogràfics diferents: el gènere (MASCULÍ/DONA) i l'edat (18-22 anys/23-45 anys/+46 anys). A més s'inclouen tres característiques demogràfiques addicionals per a cada anotador: nivell educatiu, ètnia i país de residència. [10]

text	meme	language	task1	task2
Princesa, hoy es tu día, no laves ningún plato. Gracias mi amor. BATICHOLO.COM.MX BATICHOLO BATICHOLO.COM.MX Ya mañana los lavas. f FACEBOOK.COM/BATICHOLO1	110001.jpeg	es	sexist	objectification
2+2=5 MITO Albert Einstein tenía bajo rendimiento en la escuela. VERDAD 2+2=4 CAN is El feminismo de hoy en día defiende la estupidez humana y no los derechos de las mujeres quemo ellas afirman	110006.jpeg	es	sexist	ideological-inequality
A LOS QUE NO ME SALUDAN POR EL DIA DE LA MUJER. CHUPENME LA PIJA!!! EL MAX	110009.jpeg	es	sexist	misogyny-non-sexual-violence
David Sant Hace un momento Pienso en la arrogancia que hace falta para desalmar a alguien. Unos malditos monstruos primitivos en un festín de carne, sangre y locura. #LaManada	110012.jpeg	es	sexist	sexual-violence
8 de Marzo Tawail Qu Fri "Muchas de mis películas tienen protagonistas femeninas fuertes, valientes, niñas autosuficientes que no se lo piensan dos veces antes de luchar por lo que creen con todo su corazón. Necesitarán un amigo, o un partidario, pero nunca un salvador. Cualquier mujer es capaz de ser una heroína tanto como un hombre." Hayao Miyazaki	110038.jpeg	es	non-sexist	non-sexist

[15] Prolific Academic és una plataforma en línia que connecta investigadors amb participants disposats a participar en estudis. Els participants poden trobar i completar estudis pagats sobre una varietat de temes, com psicologia, ciències socials i comportament humà.

AH! INTERNACIONAL SRMORA.COM ENTONCES SERÁ ENEL EXTRANJERO Día Internacional de la Mujer Trabajadora Desmotivación Es que existan imágenes como estas www.desmotivaciones.es	110015.jpeg	es	sexist	stereotyping-dominance

Taula 2: Exemple del conjunt de dades exist 2024

- **Seda Mut:** Aquest conjunt de dades es crea a partir de la realització d'una investigació que destaca la manaca de recursos per a l'estudi del sexisme en llengües menys representades en el processament del llenguatge natural (PLN), com el turc, i aborda aquesta bretxa en proporcionar un conjunt de dades que consta d'aproximadament 7.000 tweets i comentaris de YouTube manualment anotats.

  El conjunt de dades està creat a partir de Twiteer i Youtube, utilitzant les corresponents APIs per poder extreure'n les dades. A més, es van definir criteris d'exclusió per als tweets, com ara excloure retweets o tweets que incloguin imatges i vídeos. Cada text està classificat com a sexista o no sexista, i a més, els textos sexistes estan categoritzats a Estereotips, objectificació, misogínia, obscenitat o violència sexual, i antifeminisme (ideologia i desigualtat). On trobem cada text etiquetat en turc i anglès.

  La fiabilitat dels apunts es va verificar mitjançant el càlcul del coeficient Kappa de Cohen [25].

text	task1_en	task1_tr	task2_en	task2_tr
@yourwierdgirl bu nasil bir genelleme erkek dusmani olmussun sen haberin yok sadece	sexist	cinsiyetci	ideological-inequality	feminizm karsitligi
erkek dusmani kiz tayfa salak midir amk en sinir odlugum tayfa ya valla kafasini surte surte beynini akitican(varsa) uyuz ooldumgffe gece	sexist	cinsiyetci	misogyny-non-sexual-violence	kadin düşmanliği veya cinsel icerikli olmayan siddet
@deepwebtiktok2 bu eleman mazemeyi aldirdigindan beridir erkek dusmani oldu. kendi erkek oldugu zamanlarda yapamadiklarini suan manita olup yapilmasini bekliyor	non-sexist	hicbiri-cinsiyetci degil	non-sexist	hicbiri - cinsiyetci degil
@luganonunofkesi @begzranaj @10a10flood erkek dusmani ne alaka, sen cinsiyetcisin. yilmaz abinin bu yazilari yazmasini voleybol oynamasina baglayarak cinsiyetcilik yapıyorsun ondan sonra bize erkek dusmanj diyosun isksosksoossmd futbol oynayanlar maco, ma	sexist	cinsiyetci	objectification	objelestirme
benim sahibem erkek dusmani olmasi lazim acimamasi lazim #kole #koele #koelearıyorum #köle #finansalkoepek #finansalyardım #finansalköle #finansaldestek	sexist	cinsiyetci	sexual-violence	mustehcen veya cinsel siddet icerigi
aile kavrami bitiyor omur boyu nafaka evlilikleri ve ulkemizin gelecegini bitiriyor.erkek dusmani bir nesil yetisiyor....#tehlikeninfarkındamısınız	sexist	cinsiyetci	stereotyping-dominance	kalip dusunce / ideolojik yaklaşim veya baskinlik

Taula 3: Exemple del conjunt de dades en turc

El conjunt de dades final consta de 11.125 mostres per a l'entrenament i 7.134 mostres per a testejar, fent un total de 18.259 mostres entre els tres conjunts de dades. On els conjunts d'entrenament i testeig es seleccionen aleatòriament per garantir l'equilibri de les classes.

En primer lloc, tenim una classe que ens determina si la mostra és sexista o no sexista. I una altra classe que categoritza la mostra segons la tipologia de sexisme que s'esmenten a l'apartat 3.2. En la següent taula veiem la distribució de cadascuna de les classificacions i el volum que té cadascuna en el nostre conjunt d'entrenament:

Classe	# instàncies	% instàncies
Sexista	5511	49.54
No sexista	5614	50.46
Ideología-desigualtat	949	8.53
Estereotips-dominància	1068	9.60
Objectificació	770	6.92
Violència sexual	1359	12.21
Misogínia	1468	13.20
Total	11125	100

Taula 4: Instàncies per categoria del conjunt de dades d'entrenament

Per tant, l'estructura final del nostre conjunt de dades es la presentada a la següent taula:

ID	Language	Text	Task1	Task2
Número que identifica cada instància	Idioma en el qual esta escrita la instància	Text a classificar	Classificació del text en sexista o no sexista	Classificació del text en ideologia i desigualtat, estereotips i domini, misogínia i violència no-sexual, objectificació i violència sexual.

Taula 5: Estructura del conjunt de dades final

4.2 Processament de les dades

Per a poder entrenar els models, en primer lloc, hem de realitzar un preprocessament de les dades, és a dir, fer una preparació prèvia de les dades.

4.2.1. Neteja

En la fase de neteja de les dades, s'eliminen tots els elements no desitjats que poden afectar negativament a l'anàlisi del llenguatge. En primer lloc, eliminem els enllaços, caràcters i números no alfanumèrics. Posteriorment, transformem tot el text en minúscules, i finalment eliminem les *stopwords*.

Les *stopwords*, són paraules comunes en el llenguatge que no aporten informació específica en el contingut del text, com poden ser els articles, preposicions i conjuncions [47]. Pel que fa a l'eliminació

de les *stopwords* hem de tenir en compte l'idioma de les dades, ja que aquestes paraules són diferents en anglès, castellà i turc.

4.2.2. Tokenització

La fase de tokenització consta de dividir el text en unitats més petites anomenades *tokens*, per a facilitar el processament i anàlisi posterior. Aquesta fase es realitza de manera diferent en funció del model emprat, ja que la tokenització ha de ser coherent al model que estem utilitzant [48][49].

4.2.3. Normalització

La fase de normalització té com a objectiu estandarditzar el text per reduir la variabilitat i millorar la consistència de l'anàlisi. La normalització ens permet tractar de manera uniforme les paraules, evitant les duplicacions i simplificant el procés de l'anàlisi [50].

- ***Stemming*:** És una tècnica de normalització que consisteix a obtenir l'arrel de les paraules, eliminant els prefixos i sufixos, reduint les paraules a la seva forma base o arrel. Aquest procés ens ajuda a relacionar sota la mateixa forma totes les paraules, facilitant la identificació de patrons i extracció semàntica del text.

- **Reducció de les paraules al seu lema:** La reducció de paraules al seu lema és similar a *stemming*, però en lloc d'eliminar els prefixos i sufixos, redueix els lemes utilitzant un diccionari lèxic. Això ens permet obtenir una representació més precisa de la paraula.

4.2.4. Vectorització

La vectorització és el procés de convertir el text en una representació numèrica. On cada *token* es representa com un vector numèric, i la dimensió del vector representa una característica del *token*.

En funció del model, hem utilitzat diferents tècniques de vectorització:

- **TF:** La vectorització TF (freqüència de terme) mesura la freqüència d'una paraula en el document. No considera la importància relativa, sinó que calcula la proporció de la quantitat de cops que apareix una paraula en relació amb tot el conjunt de paraules.

  On t és la paraula i d el document, es calcula:

$$TF\ (t,d) = \frac{\#\ de\ vegades\ qeu\ t\ apareix\ en\ d}{\#\ de\ termes\ totals\ en\ d}$$

- **TF-IDF:** La vectorització TF-IDF combina TF amb IDF per avaluar la rellevància de cada paraula. On IDF mesura quan de rara o comú és la paraula en tot el conjunt de textos (*tweets*). Per tant, TF-IDF pondera les paraules segons la seva freqüència en el text i la seva raresa en el conjunt de textos.

  Es calcula:

$$IDF\ (t,D) = \log\left(\frac{N}{DF(t,D)}\right)$$

$$\mathrm{TF-IDF(t,d,D)} = \mathrm{TF(t,d)} \times \mathrm{IDF(t,D)}$$

  On t és la paraula, D el conjunt de documents (*corpus*), N el nombre total de documents i DF(t,D) és el nombre de documents que contenen el terme t.

- **Word Embeddings:** *Word embeddings* són representacions numèriques que capturen el significat i context de les paraules utilitzant métodes com *Word2Vec*[16] o *FastText*[17]. En aquest cas, es representen les paraules o *tokens* en un espai vectorial que està basat en el context semàntic. Aquesta tècnica de normalització és la que utilitzem per als models de PLN [51].

4.3 Llibreries

En el procés de neteja de les dades, anàlisis i entrenament dels models, s'han utilitzat diverses llibreries de Python:

- Pandas: proporciona estructures de dades i eines per a l'anàlisi. S'utilitza per a manipular i analitzar.
- Matplotlib: s'utilitza per a la visualització de les dades i crear gràfics.
- Google.colab: és un entorn de notebook basat en el núvol que ens permet executar codi Python, llegir i escriure en fitxers emmagatzemats a Google Drive.
- Ast: Proporciona funcions per a treballar amb arbres, permetent analitzar i manipular el codi
- Re: proporciona funcions per a treballar amb expressions regulars o patrons, que ens permeten buscar i manipular cadenes de text.
- Snowballstemmer: l'utilitzem per a normalitzar les paraules.
- Sklearn: és una biblioteca de codi obert per a l'aprenentatge automàtic. Proporciona una gran varietat d'algorismes i eines per avaluar i ajustar els models.
- Numpy: s'utilitza per a la manipulació de matrius i vectors.
- Plotly: Es fa servir per generar gràfics a partir de les dades. És una biblioteca de visualització de dades a Python que permet crear gràfics interactius i dinàmics.
- JSON: S'utilitza per treballar amb dades en format JSON (JavaScript Object Notation).
- Datetime: Mòdul per manejar dates i temps.
- openpyxl: S'utilitzat per treballar amb fitxers Excel (treballem amb excel per a importar els resultats dels models).
- load_workbook: Per carregar fitxers Excel.
- Transformers: biblioteques de Hugging Face per treballar amb models de transformació, com BERT, RoBERTa, i XLM-RoBERTa.
- Datasets: Llibreria per gestionar conjunts de dades, especialment en el context d'aprenentatge automàtic.
- PyArrow: S'utilitza per manipular les dades en memòria i un emmagatzematge eficient, especialment en grans volums de dades.
- PyTorch: biblioteca d'aprenentatge profund per construir i entrenar models neuronals.
- Hugging Face Trainer: API que facilitar l'entrenament de models de transformació emprats en el treball.

4.4 Eines i tècniques PLN

Per al desenvolupament del treball s'han utilitzat les eines següents:

- **Google Colaboratory**: Colaboratory és una plataforma gratuïta basada en el núvol que permet executar codi Python a Jupyter Notebooks sense necessitat d'instal·lar cap programa local. La funcionalitat principal és la creació i edició de codi per a la realització del treball, així com l'execució del codi i la visualització dels resultats [52].

[16] *Word2Vec* és una tècnica per representar paraules com a vectors de nombres reals. Això vol dir que a cada paraula se li assigna una seqüència única de números que captura el seu significat i les relacions semàntiques amb altres paraules.
[17] *FastText* és una biblioteca de codi obert per aprendre i avaluar incrustacions de paraules i models de classificació de textos. Va ser creat pel laboratori d'AI Research (FAIR) de Facebook.

- **Jupyter Notebook**: Jupyter Notebook és un entorn interactiu per al desenvolupament de codi i la visualització de dades. Es pot combinar codi, text, *markdown* i contingut multimèdia en un únic document, el que facilita la creació i seguiment del treball. La funcionalitat principal ha sigut l'escriptura de codi intercalada amb explicacions de seguiment, així com poder executar cel·la a cel·la.

- **Visual Studio Code (VS Code)**: És un editor de codi font altament versàtil que va ser desenvolupat per Microsoft. Té una àmplia gamma d'extensions, la qual cosa ens ha permès desenvolupar la web interactiva utilitzant streamlit.

- **HCP de la Universitat**: L'HCP (*High Performance Computing*) de la Universitat Pompeu Fabra és un servei que proporciona accés a recursos informàtics d'alt rendiment per a la recerca i la docència. Ens permet executar codi en nodes amb més potència de processament i memòria que els ordinadors personals. La funcionalitat implementada ha sigut per a l'execució de codi amb jupyter notebook i l'emmagatzematge dels resultats obtinguts en cadascuna de les diferents execucions [53].

- **Putty**: És un terminal client de codi obert que facilita la connexió a servidors remots mitjançant SSH, Telnet i altres protocols. En aquest cas, s'empra com a eina per a realitzar la connexió des del port del servidor al port físic del nostre ordinador. Així com per executar jupyter notebook en el servidor.

- **FortiClient VPN**: Proporciona un accés segur a les xarxes privades des d'ubicacions remotes. La funcionalitat principal en aquest treball ha sigut per a tindre una connexió segura a la xarxa universitària i tenir accés remot als recursos.

- **GitHub**: Plataforma online per allotjar repositoris Git i compartir codi amb altres usuaris.

La combinació d'aquestes eines ha permès desenvolupar aquest treball de fi de grau de manera eficient i col·laborativa.

4.5 Models de classificació

En aquest apartat, analitzarem de manera detallada els diferents models de classificació utilitzats en aquest treball. Un model de classificació és un algorisme d'aprenentatge automàtic que s'utilitza per assignar etiquetes a les dades d'entrada basant-se en les seves característiques. L'objectiu principal d'un model de classificació és predir la categoria o classe a la qual pertany una nova observació, a partir d'un conjunt de dades d'entrenament previ que conté exemples amb etiquetes conegudes. Hem emprat models clàssics d'aprenentatge automàtic com *Multinomial Naive Bayes*, *K-Nearest Neighbors* (KNN), *Regressió Logística*, *Random Forest* i *Gradient Boosting*. Tot i que aquests models són simples, ens permeten examinar el comportament del llenguatge i les expressions sexistes. L'objectiu és avaluar el rendiment de cadascun dels models en la classificació i categorització de textos sexistes.

4.5.1. Multinomial Naive Bayes (MNB)

El model *Multinomial Naive Bayes* és un classificador probabilístic basat en el teorema de Bayes[18] sobre la independència condicional de les característiques.

Aquest model calcula la probabilitat que el text pertany a una de les classes. En el nostre cas, per a la primera tasca, calcula la probabilitat que el text que sigui sexista o no sexista. I en la segona tasca, calcula la probabilitat de pertànyer a cadascuna de les tipologies descrites anteriorment [55].

[18] Expressió matemática: P(A|B) = (P(B|A) * P(A)) / P(B) [54]

4.5.2. K-Nearest Neighbors (KNN)

L'algoritme de *K-Nearest Neighbors* classifica les dades basant-se en les classes dels nodes veïns. Mesura quan de properes són les dades amb una mètrica de distància, com podria ser la distància euclidiana.

El valor K ens indica el nombre de veïns més propers que s'estan considerant, assignant al text que s'està classificant la classe que predomina entre els k veïns.

Per poder alimentar aquest algoritme necessitem representar el text de manera numèrica, utilitzant el model "*bag of words*" amb tf-idf (freqüència de terme–freqüència inversa de document) que veiem explicat a l'apartat de vectorització 4.2.4. [56][57]

4.5.3. Random Forest

El *Random Forest* és un conjunt d'arbres de decisió que es combinen per assolir prediccions més robustes i precises. Cada arbre al bosc s'entrena amb una mostra aleatòria de les dades i utilitza un subconjunt aleatori de característiques per prendre decisions. Posteriorment, les prediccions de tots els arbres es combinen per obtenir la classificació [58] [58].

Igual que per a KNN, devem representar el text de manera numèrica, per a poder entrenar el model amb vectors i etiquetes de classificació.

4.5.4. Gradient Boosting

Gradient Boosting és una tècnica d'aprenentatge automàtic que construeix un model predictiu a partir de petites prediccions amb arbres de decisió. A diferència del *random forest*, on els arbres es construeixen de manera independent, en el *gradient boosting* es construeixen de manera seqüencial, ajustant els errors dels arbres anteriors [59].

En el cas de la primera tasca, on només es realitza una classificació binària, s'introdueix un sol arbre de decisió.

En aquest model es tenen en compte diversos paràmetres que ajuden a obtenir un millor resultat:

- *Learning rate*: La taxa d'aprenentatge controla com cada arbre contribueix en el model, de manera que si establim un valor petit, reduïm la influència dels diferents arbres, i si establim un valor elevat, als arbres influiran d'una manera més significativa.
- *Estimators*: És el que coneixem com nombre d'etapes a realitzar pel model. És un valor important per a ajustar les prediccions, però cal trobar l'equilibri per no sobre ajustar el model.
- *Subsample*: Determina la fracció de mostres utilitzades per ajustar els arbres individuals. Un valor menor introdueix més biaix, però redueix la variància.

4.5.5. Regressió Logística

La regressió logística és un mètode estadístic utilitzat per predir la probabilitat que una observació pertanyi a una de diverses categories possibles. S'utilitza en diversos camps per a tasques de classificació binària i multiclasse [60]. En el context d'aquest treball, la regressió logística s'utilitza per a la classificació de textos sexistes.

En el nostre cas s'utilitza per a la classificació binària de si un text és sexista o no sexista i per la classificació en cadascuna de les tipologies d'aquells textos que són sexistes.

El model calcula la probabilitat de classificació del text en cadascuna de les classes i assigna l'etiqueta amb la probabilitat més alta. Aquesta probabilitat es calcula utilitzant una funció sigmoide que

transforma una combinació lineal de les característiques d'entrada en un valor entre 0 i 1, representant la probabilitat que la instància pertanyi a una de les classes [61].

4.6 Models de PLN

El processament de llenguatge natural (PLN) és una branca de la intel·ligència artificial que s'ocupa de proporcionar a les computadores la capacitat de comprendre textos i paraules de la mateixa manera que ho faria un humà. Ja que permet reconèixer, interpretar i manipular el llenguatge humà [62].

En aquest apartat veurem de manera detallada els diversos models de PLN utilitzats en el treball, que ens permeten una comprensió més profunda i contextualitzada dels textos en diversos idiomes.

Els models preentrenats que s'utilitzen són models basats en arquitectura *Transformer*. L'arquitectura *Transformer* està dissenyada per a manipular dades seqüencials, com el llenguatge natural, de manera més eficient i efectiva. Aquesta arquitectura ha revolucionat el camp del processament de llenguatge natural (PLN) i s'ha convertit en la base de molts models. Aquesta arquitectura consta de dues parts principals el codificador (*encoder*) i el descodificador (*decoder*). Encara que en els models implementats en aquest treball només es fa servir la part del codificador [63].

Cada codificador està compost per múltiples capes idèntiques, que inclouen les següents subcapes:

- Mecanisme de *Self-Attention*: Permet que el model avaluï la importància de cada paraula (o token) del text en relació amb totes les altres paraules. Es calculen pesos que determinen quines paraules són més rellevants.
- Capa *Feedforward*: Una xarxa neuronal *feedforward* totalment connectada que s'aplica de manera independent a cada posició.

Cadascuna d'aquestes subcapes està envoltada per una capa de normalització i un mecanisme de connexions residuals per facilitar el flux de gradients durant l'entrenament [64][64].

4.6.1. BERT (*Bidirectional Encoder Representations from Transformers*)

El model de BERT, és un model de llenguatge preentrenat desenvolupat per *Google AI Language* l'any 2018 impulsat per a desenvolupar tasques de NPL gràcies a la capacitat de processar i comprendre el llenguatge natural d'una manera més profunda que els models simples d'aprenentatge automàtic. Aquest model es basa en una arquitectura de transformadors bidireccionals, capaços de capturar el context de les paraules dintre d'un text.

Durant la fase d'entrenament, BERT utilitza el model de llenguatge emmascarat (MLM). MLM és una tècnica que emmascara aleatòriament paraules de l'oració i s'entrena el model per a predir aquestes paraules en funció del context. És una tècnica que ajuda el model a entendre i generar text de manera més efectiva segons el context. Això permet a BERT captar millor els matisos i les ambigüitats del llenguatge humà [65][66].

En aquest treball també s'utilitza el model BERT Multilingual (*bert-base-multilingual-cased*), que és una versió preentrenada en 104 idiomes, incloent-hi l'anglès, el castellà i el turc.

A continuació veiem les versions i models emprats:

- **bert-base-uncased:** És el model BERT original preentrenat per Google. Està preentrenat en anglès [67].
- **dccuchile/bert-base-spanish-wwm-uncased:** És una variant del model BERT entrenat específicament per al castellà. "WWM" significa "Whole Word Masking", és una tècnica que emmascara paraules completes en lloc de subparaules durant el preentrenament [68].
- **dbmdz/bert-base-turkish-cased:** És un model BERT preentrenat per a l'idioma turc. "Cased" significa que el model diferencia entre majúscules i minúscules [69].

- **bert-base-multilingual-cased:** És un model BERT multilingüe que està entrenat en més de 100 idiomes [70].

2.6.2. RoBERTa (*Robustly optimized BERT approach*)

El model de RoBERTa és una variant del model BERT que s'entrena a partir de tècniques de preentrenament més avançades i un conjunt de dades molt més gran. Va ser desenvolupat per Facebook AI Research en 2019.

A diferència del model de BERT, aquest model inclou fonts addicionals com *Common Crawl*, Wikipedia i llibres, entre d'altres. A més, RoBERTa s'entrena amb més iteracions i major tamany del *batch*, permetent un major aprenentatge [71].

El model de RoBERTa també utilitza *Masked Language Modeling* (MLM) en la fase d'entrenament. A diferència de BERT, que utilitza un enfocament d'emmascarament de *tokens* estàtic, RoBERTa utilitza un emmascarament dinàmic. El que significa que la proporció de *tokens* que s'emmascara es variable, fent que el model aprengui dependències més robustes entre les paraules. A més, aquest model és capaç de predir una oració completa.

Gràcies a aquestes modificacions, RoBERTa ha demostrat un millor rendiment en diverses tasques de processament del llenguatge natural i en *benchmarks* estàndard com *GLUE*, *SQuAD* i *RACE*, superant BERT en diversos [72][73].

De la mateixa manera, en aquest treball utilitzem també el model XLM-RoBERTa (Cross-lingual Language Model RoBERTa). Aquest model s'entrena amb un conjunt de dades en més de 100 idiomes, la qual cosa que permet aprendre patrons i relacions lingüístiques en una àmplia gamma de llengües. A més, inclou la traducció automàtica i la classificació de textos per idioma, cosa que obliga el model a desenvolupar una comprensió més profunda de les relacions entre els idiomes.

A continuació veiem les versions i models emprats:
- **roberta-base:** És el model RoBERTa, una variant millorada de BERT [74].
- **PlanTL-GOB-ES/roberta-base-bne**: És un model RoBERTa ajustat específicament per al castellà, desenvolupat pel PlanTL (Pla de Tecnologies del Llenguatge) del Govern d'Espanya. Fes servir dades de la Biblioteca Nacional d'Espanya (BNE) [75].
- **xlm-roberta-base:** És una variant multilingüe de RoBERTa, dissenyada per manejar múltiples idiomes de manera robusta. Està entrenat en 100 idiomes [76].

4.7 Entrenament dels models aplicats

En l'apartat d'entrenament es descriu el procés d'entrenament dels models implementats i explicats anteriorment.

4.7.1 Models de classificació

En quant als models clàssics d'aprenentatge automàtic s'ha emprat la tècnica de *pipeline*[19].

Cadascún dels models s'han entrenat utilitzant 4 conjunts de dades diferents. Tres d'ells format per les dades en l'idioma corresponent (anglés, castellà i turc) i l'altre es la unió d'aquests tres, obtenint així un conjunt de dades multilingüe de tres idiomes. Aquest enfocament multilingüe ens permet avaluar la robustesa i el rendiment de cada model en diferents contextos lingüístics.

[19] Un pipeline d'aprenentatge automàtic és una seqüència de passos interconnectats que automatitzen el procés de creació, formació, avaluació i desplegament de models d'aprenentatge automàtic [77].

Veiem el procés detallat de la configuració dels *pipelines* emprats en cadascun dels models en la següent figura:

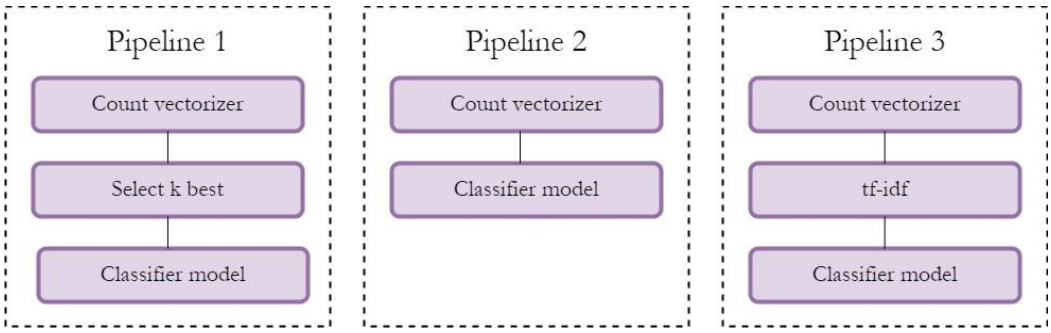


Figura 1: Esquema pipelines

En tots tres *pipelines* apliquem, en primer lloc, mètode de vectorització *CountVectorizer* (secció 4.2.4) per a transformar el text en un vector numèric. En el cas del pipeline 1, apliquem *SelectKBest*[20], on s'han testejat diferents amb diferent nombre de mostres (valors de k) i en cas de cada model el valor escollit és diferent per a trobar el millor resultat. En el cas del pipeline 3, utilitzem TfidfTransformer abans de l'algoritme de classificació.

Un cop aplicats els mètodes vectorització o selecció de característiques, apliquem els diferents models utilitzats: *Multinomial Naive Bayes*, *K-Nearest Neighbors* (KNN), *Regressió Logística*, *Random Forest* i *Gradient Boosting*.

4.7.2 Models preentrenats de PLN

En el cas dels models preentrenats, la part més rellevant implementada es la fase de *fine-tunning*. Per a realitzar el *fine-tunning*[21], és necessari tenir un conjunt d'entrenament i un de validació per avaluar el rendiment. És per això que tenim dividit el nostre conjunt de dades de manera aleatòria en aquests dos conjunts. On el conjunt d'entrenament representa el 80% i el conjunt de validació el 20% restant. A continuació, carreguem el model corresponent i realitzem un ajustament dels paràmetres. Els paràmetres emprats són els següents:

- *output_dir*: especifica el directori on es guarden els resultats de l'entrenament.
- *evaluation_strategy*: especifica l'estratègia d'avaluació del model. En aquest cas utilitzem epoch, cosa que s'avalua el model al final de cada època.
- *logging_strategy*: especifica l'estratègia de registre de logs. Utilitzem steps, per tant, registren els logs cada cert nombre de passos en lloc de cada època, permetent un monitoratge més freqüent.
- *save_strategy*: S'especifica com es desa el model. En aquest cas, es guarda el model al final de cada època.
- *learning_rate*: La taxa d'aprenentatge utilitzada per a l'optimitzador durant l'entrenament és de 3e-5.
- *per_device_train_batch_size*: Mida del bloc d'entrenament, en aquest cas se n'utilitza 16.
- *per_device_eval_batch_size*: Mida del bloc d'avaluació per dispositiu.

[20] SelectKBest és un algorisme de selecció de característiques que identifica les característiques més rellevants per a un model d'aprenentatge automàtic. És un algorisme d'aprenentatge supervisat, és a dir, requereix dades etiquetades per entrenar. L'algorisme funciona avaluant cada característica en funció de la seva capacitat per distingir entre les diferents classes de les dades.

[21] La fase d'ajustament d'un model d'aprenentatge automàtic és el procés d'ajustar els paràmetres del model per millorar-ne el rendiment en una tasca o conjunt de dades específics.

- *num_train_epochs*: Determineu el nombre d'èpoques que s'executarà l'entrenament. En el nostre cas, utilitzem 10 èpoques
- *weight_decay*: La taxa de decaïment dels pesos per a la regularització del model, que ajuda a prevenir el sobreajustament. Utilitzem la taxa de 0.1
- *logging_steps*: Determina la freqüència de registre de logs en termes de nombre de passos. Aquí es registren els logs cada 100 passes.
- *load_best_model_at_end*: Serveix per indicar que cal carregar el millor model al final de l'entrenament, ja que s'estableix TRUE.
- *metric_for_best_model*: Determina la mètrica utilitzada per determinar quin és el millor model. En aquest cas, es fa servir la precisió (accuracy).
- *gradient_accumulation_steps*: Determina el nombre de passos d'acumulació de gradients. Augmentar la mida del lot d'entrenament sense utilitzar més memòria.

Un cop carregat el model, instanciem una instància de *Callback*[22], que encarregarà de guardar les mètriques de l'entrenament. També es crea una instància de la classe *Trainer* on especifiquem el model a entrenar, els arguments, el conjunt de dades a utilitzar per entrenar el model, el conjunt de dades de validació i la llista per a guardar les mètriques. I ja procedim a entrenar el model. Un cop el model està entrenat, s'avalua amb el conjunt de dades de validació. Finalment, guardem els resultats [78].

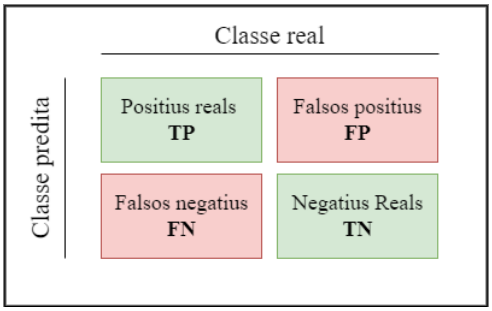
4.8 Avaluació dels models

En aquest apartat veiem les mètriques d'avaluació utilitzades per a analitzar el rendiment dels models utilitzar en l'estudi. L'avaluació dels models pren un paper rellevant en la comprensió dels resultats obtinguts per cada model.

Aquestes mètriques ens permeten valorar l'eficàcia dels models en la detecció del llenguatge sexista a les xarxes socials. Ja que cada mètrica ens proporciona un resultat que indica la capacitat que té el model per identificar correctament la classificació de les variables en les diferents tasques.

Abans de veure en profunditat les mètriques, introduïm el concepte de matriu de confusió. És una eina fonamental per a avaluar el rendiment dels models, ja que representa una taula amb les prediccions correctes o incorrectes realitzades per al model.

Veiem l'estructura de la matriu i el significat de cadascun dels termes:



Positius reals (TP): Casos on el model prediu correctament la classe positiva.
Falsos Positius (FP): Casos on el model prediu incorrectament la classe positiva.
Falsos Negatius (FN): Casos on el model prediu incorrectament la classe negativa.
Negatius reals (TN): Casos on el model prediu correctament la classe negativa.

Figura 2: Matriu de confusió

[22] Una instància de callback es refereix a una funció que es passa com a argument a una altra funció, la qual la invoca (o "truca de tornada") en un moment específic durant la seva execució. Aquest mecanisme és fonamental per a la programació asíncrona i esdeveniments, permetent que una funció notifiqui a una altra quan s'ha completat una tasca o quan s'ha produït un esdeveniment específic.

4.8.1. Accuracy

La precisió o *accuracy*, mesura la freqüència en què un model prediu correctament el resultat. El valor obtingut és resultat de dividir el nombre de prediccions correctes entre el total de totes les prediccions. Per tant, mesura la proporció d'instàncies classificades correctament [79].

$$Accuracy = \frac{\#instàncies\ classificades\ correctament}{\#\ instàncies\ total} = \frac{TP + TN}{TP + TN + FP + FN}$$

Pot prendre un valor comprès entre 0 i 1, on 1 significa que totes les instàncies s'han classificat correctament, mentre que 0 indica que cap predicció és correcta.

4.8.2. Precision

La precisió mesura la proporció de prediccions positives correctes realitzades pel model. Es calcula dividint el nombre d'instàncies correctament classificades (prediccions positives correctes) entre el nombre total d'instàncies classificades com a positives. Per tant, ens proporciona una mesura de la precisió de classificació del model.

$$Precision = \frac{TP}{TP + TN}$$

Els valors de precisió estan en el rang de 0 a 1, on 1 representa una precisió perfecta, cosa que significa que totes les prediccions positives són correctes, mentre que 0 indica una manca total de precisió, cosa que implica que cap predicció positiva és correcta . La precisió és una mètrica crucial per avaluar la capacitat del model per fer prediccions precises d'una classe específica en un problema de classificació [79].

4.8.3. Recall

El *recall* indica quina proporció de les instàncies positives reals han sigut detectades correctament pel model. Per tant, es calcula dividint el nombre de veritables positius (instàncies correctament classificades com a positives) entre el nombre total d'instàncies que realment són positives a les dades.

$$Recall = \frac{TP}{TP + FN}$$

El *recall* pot prendre valors en el rang de 0 a 1, on 1 representa una exhaustivitat perfecta, és a dir, totes les instàncies positives reals van ser identificades correctament pel model, mentre que 0 indica una manca total d'exhaustivitat, cosa que implica que cap instància positiva real va ser detectada pel model [79].

4.8.4. F1 score

La mètrica *F1-score* és una mètrica utilitzada per avaluar el rendiment del model, combinant la precisió i el recall en una sola mètrica. El *F1-score* es defineix com la mitjana harmònica de la precisió i el *recall*.

$$F1\ score = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

El *F1-Score* pot prendre valors al rang de 0 a 1, on 1 representa el millor rendiment possible, és a dir, tant la precisió com el recall són perfectes. Un valor de 0 indica el pitjor rendiment possible [79].

4.8.5. Loss

La mètrica de pèrdua o *loss* representa la diferència entre les prediccions del model i els valors reals. Existeixen diferents funcions de pèrdua com: Pèrdua d'entropia creuada (*Cross-Entropy Loss*), error quadràtic mitjà (*Mean Squared Error*) i error absolut mitjà (*Mean Absolute Error*).

En aquest treball, s'utilitza únicament *Cross-Entropy Loss*, ja que estem analitzant el rendiment d'un problema de classificació. Aquesta mètrica mesura la discrepància entre les distribucions de probabilitat predites i les classes reals.

En el cas de la classificació binària per a la tasca 1, on y és l'etiqueta de classificació real (0 no sexista, 1 sexista) i $\hat{y}$ és la probabilitat predita per al model de què la mostra pertanyi a la classe 1:

$$\text{Cross} - \text{Entropy Loss} = -(ylog(\hat{y}) + (1-y)\log(1-\hat{y}))$$

En el cas de la classificació de la tasca 2, on n representa el nombre de classes:

$$\text{Cross} - \text{Entropy Loss} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

Contra menor sigui el valor resultant, millor és el rendiment del model. Perquè un valor baix indica que les prediccions realitzades pel model s'assemblen a les etiquetes reals [79].

Capítol 5

Anàlisi i resultats

En aquest apartat trobem les dades resultants del treball realitzat, això com les limitacions presentades. Els resultats es divideixen en diferents seccions, on es reflecteixen els resultats obtinguts a les diferents classificacions realitzades pels diferents models exposats.

5.1 Interpretació dels resultats

En aquest apartat es recullen els resultats obtinguts en les diferents tasques de classificació per als models implementats. En primer lloc, trobem els resultats obtinguts en realitzar la classificació binària dels comentaris en sexistes o no sexistes. En segon lloc, trobem els resultats obtinguts a partir de realitzar la categorització dels comentaris en les diferents tipologies esmentades al llarg del treball. Finalment, trobem els resultats obtinguts en la classificació i categorització dels models preentrenats.

5.1.1 Models de classificació

En aquesta secció es presenten els resultats obtinguts de la classificació i categorització dels comentaris a les xarxes socials utilitzant els diferents algorismes d'aprenentatge automàtic (models de classificació) que s'han esmentat anteriorment. Els models avaluats són: *Multinomial Naive Bayes, K-Nearest Neighbors (KNN), Regressió Logística, Random Forest* i *Gradient Boosting.*

Cada model s'ha entrenat per separat a cadascun dels idiomes (anglès, castellà i turc) i després amb el conjunt de dades que conté els tres idiomes conjuntament. Per a cada model, les mètriques que s'han fet servir per avaluar el rendiment són la precisió, exactitud, recall i F1-score.

Per al model *Naive Bayes* s'observa que el rendiment varia significativament entre els diferents idiomes i les pipelines. En general, el tercer *pipeline* ofereix el millor rendiment a totes les mètriques, particularment per a l'idioma turc i la combinació de tots els idiomes. El segon *pipeline* també mostra un rendiment sòlid, destacant els resultats obtinguts per al turc. S'observa que el resultat obtingut en el primer *pipeline* indiquen un rendiment pitjor per a tots els idiomes, destacant el baix rendiment de la combinació de tots els idiomes. Per tant, els *pipelines* 2 i 3 són més efectius per a la classificació binària de comentaris a la xarxa socials utilitzant el model *Multinomial Naive Bayes* (veure Taula 6).

En el cas dels resultats obtinguts per a la categorització dels textos, s'observa un rendiment inferior en tots els *pipelines*, encara que per a primer *pipeline* el rendiment és lleugerament superior per a l'anglès i el castellà. En canvi, per al turc, el segon *pipeline* presenta un millor rendiment el segon *pipeline*. En el cas dels resultats obtinguts a partir d'entrenar el model amb tots els idiomes conjuntament, s'observa un baix rendiment, amb un *acuracy* mitjà de 0,54 (veure Taula 7).

L'anàlisi dels resultats mostra que el rendiment del model *K-Nearest Neighbors* (KNN) també varia considerablement entre els diferents idiomes i *pipelines*. En aquest cas, el primer *pipeline* ofereix el millor rendiment per als idiomes individuals, destacant que per al turc obté un millor rendiment. No obstant això, la combinació de tots els idiomes presenta un desafiament significatiu a tots els *pipelines*, amb el tercer *pipeline* mostrant una millora notable en precisió però amb una disminució significativa en el *F1 score*. Els resultats obtinguts suggereixen que el model KNN té dificultats per a realitzar la classificació quan es combinen tots els idiomes, i el rendiment és més consistent quan s'enfoca en un sol idioma. A més, el tercer *pipeline* sembla gestionar millor la combinació d'idiomes en termes de precisió, tot i que encara hi ha marge de millora en termes de *recall* i *F1 score* (veure Taula 8).

En canvi, els resultats obtinguts en la categorització dels textos segons la tipologia de sexisme, el rendiment del model és molt baix. No hi ha gaires diferències entre els diferents idiomes i *pipelines*. El model de KNN ha obtingut el pitjor resultat en l'entrenament dels models. Concretament en el primer *pipeline* per al conjunt de dades que conté tots els idiomes (veure Taula 9).

En el cas del model *Random Forest*, també varia entre els diferents idiomes i *pipelines*, però en general ofereix un rendiment més robust i consistent per a la classificació dels textos sexistes. El primer *pipeline* presenta un rendiment moderat amb l'idioma turc que mostra els millors resultats. El segon *pipeline* millora les mètriques generals, especialment per a l'idioma turc i la combinació de tots els idiomes. El tercer *pipeline* manté un rendiment estable, semblant al segon *pipeline*. Per tant, el model presenta un rendiment sòlid a la classificació binària de comentaris en xarxes socials, amb un rendiment particularment més acurat en l'idioma turc, encara que té una bona capacitat per manipular múltiples idiomes en un sol model. En termes de *pipelines*, el segon i tercer *pipeline* ofereixen els millors rendiments generals (veure Taula 10).

En el cas dels resultats obtinguts en la categorització, veiem que el rendiment del model també varia significativament entre els diferents idiomes. Per a tots els *pipelines*, l'idioma turc presenta els millors resultats en termes *d'accuracy*, *precision*, *recall* i *F1 score*, mentre que l'anglès té els pitjors resultats. Entre el segon i tercer *pipeline* no s'observen gaires diferències de les mètriques avaluades, i aquestes són millors a les obtingudes en el primer pipeline. Destaquem una correspondència uniforme entre la *precision* i el *recall* en cada idioma dins de cada *pipeline*. Això suggereix que el model té un equilibri entre la capacitat d'identificar correctament els casos positius i els negatius, encara que els resultats no són gaire acurats (veure Taula 11).

En el cas dels resultats obtinguts a l'entrenament del model *Gradient boosting*, destaquem el rendiment en la classificació dels comentaris sexistes obtingut del model en els diferents *pipelines* amb l'idioma turc. Encara que els resultats també varien en funció de l'idioma del conjunt de dades, s'observa una major estabilitat del model en tots els idiomes (veure Taula 12).

En canvi, el rendiment del model en la categorització dels comentaris obté millors resultats per al castellà, on els resultats en els tres *pipelines* és prou estable i similar. El model *Gradient boosting* obté millors resultats que els anteriors models en la tasca de categorització, però no són molt positius (veure Taula 13).

En el cas del model de regressió logística, en general es mostra un rendiment sòlid i estable per als diferents idiomes. En aquest cas, també destaquem el rendiment dels models entrenats amb el conjunt de dades turc, sobretot el *pipeline* 3 aconsegueix tenir un *Recall* de 0.9 al tercer pipeline per a la classificació binària, el millor resultat de tots els models. S'observa que el segon *pipeline* mostra petites millores en el rendiment general per al conjunt de dades que conté tots els idiomes comparat amb *pipeline* 1 (veure Taula 14).

Si observem els resultats obtinguts per a la categorització dels comentaris, el turc continua sent el conjunt de dades que millor rendiment té. No hi ha diferències significatives en el rendiment del segon i tercer *pipeline* per al turc, però sí per a la resta d'idiomes. El primer *pipeline* mostra uns resultats menys precisos per a totes les mètriques avaluades (veure Taula 15).

En conclusió, els resultats obtinguts a partir de l'entrenament dels diferents models suggereixen que l'elecció del *pipeline* i l'idioma tenen un impacte significatiu en el rendiment dels models. I els models tenen un major rendiment en la tasca de classificació dels comentaris sexistes, però un rendiment inferior a l'hora de dur a terme la tasca de categorització.

5.1.2 Models preentrenats de PLN

En aquesta secció es presenten els resultats obtinguts a partir de la execució dels models preentrenats BERT i RoBERTa en les diferents versions monolingües i multilingües. En primer lloc, trobem els

resultats obtinguts alhora de classificar els textos en sexistes o no sexistes (tasca 1), i en segon lloc trobem els resultats de la categorització dels textos sexistes (tasca 2).

- **BERT**: En el cas del model BERT tal com veiem a la secció 4.6 Models de PN, s'ha entrenat primerament amb versions específiques per cada idioma (anglès, castellà i turc) i posteriorment amb una versió multilingüe per a cada idioma per separat i el conjunt de dades que conté totes elles.

  A les següents gràfiques es reflecteixen els resultats obtinguts de la mètrica *loss* en cada *epoch*, avaluada a les dades d'entrenament i validació.

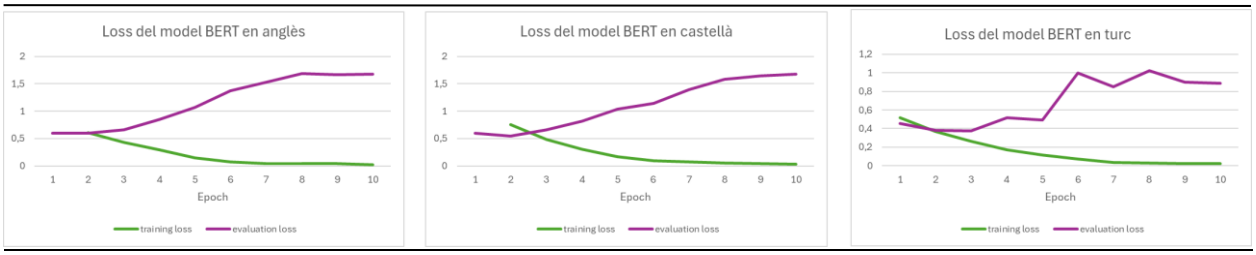


Figura 3: Resultats de la mètrica loss per a la tasca 1 en el model BERT emprant versions específiques per idioma

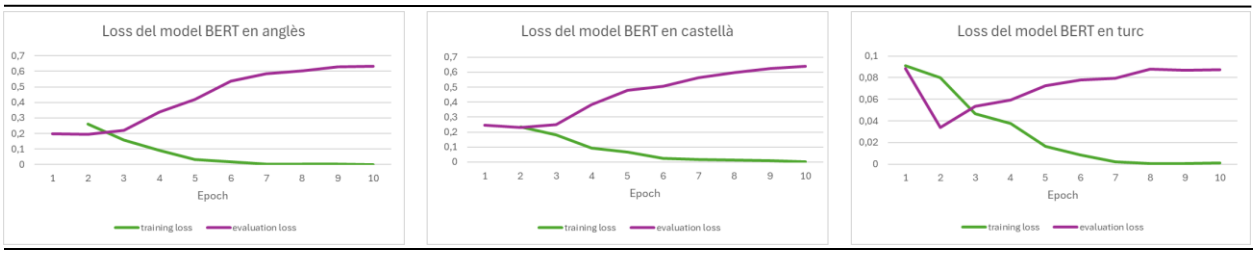


Figura 4: Resultats de la mètrica loss per a la tasca 2 en el model BERT emprant versions específiques per idioma

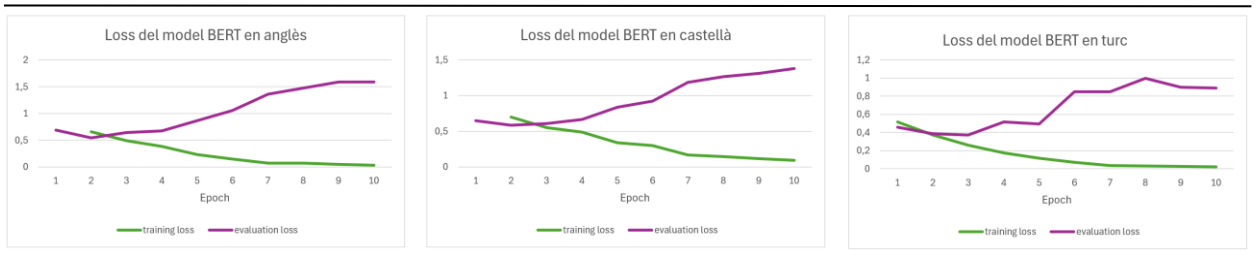


Figura 5: Resultats de la mètrica loss per a la tasca 1 en el model BERT emprant la versió multilingüe

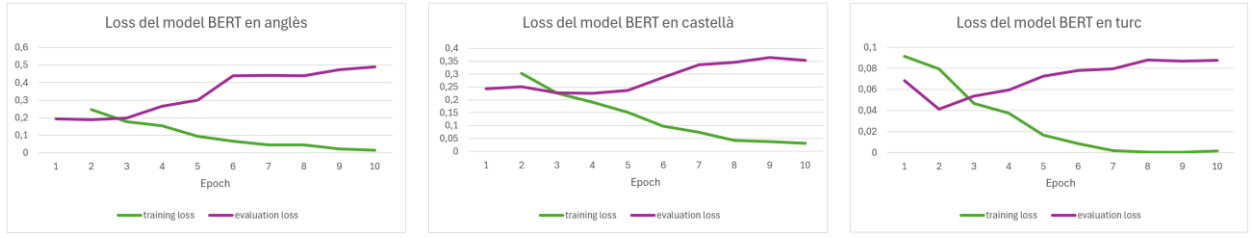


Figura 6: Resultats de la mètrica loss per a la tasca 2 en el model BERT emprant la versió multilingüe

Analitzant els resultats mostrats en les figures anteriors, veiem que la funció *loss* en els diversos models es comporta de manera similar. On el valor *loss* del conjunt d'entrenament (*training loss*) disminueix en cada *epoch* arribant a ser molt proper a 0. Però, en canvi, el comportament d'aquesta mètrica en el conjunt de dades d'avaluació (*evaluation loss*) és contrari, comença sent decreixent o similar en les primeres *epoch*, però a partir del segon o tercer *epoch* comença a ser creixent. El qual ens indica que el nostre model està sobre ajustat, és a dir, té *overfitting*.

Si entrenem el model BERT amb el conjunt de dades que conté els tres idiomes (veure Figura 7: Resultats de la mètrica loss en el model BERT emprant la versió multilingüe i el conjunt de dades de tots els idiomes), el comportament de la mètrica loss és similar.

Figura 7: Resultats de la mètrica loss en el model BERT emprant la versió multilingüe i el conjunt de dades de tots els idiomes

La Figura 8: Resultats de la métrica accuracy per al model BERT emprant versions específiques per idioma correspon als resultats del *accuracy* obtinguts en el model BERT entrenat amb la versió específica per a cada idioma en les dues tasques propostes. On veiem un comportament prou lineal en tots els idiomes. Destaquem que en la tasca de categoritzar els textos en sexistes o no sexistes (tasca 1) el model entrenat amb l'idioma turc obté els millors resultats, amb un *accuracy* mitjà de 0,82. Però en la tasca de categoritzar els textos sexistes el rendiment és molt inferior, arribant a tenir un *accuracy* mitjà inferior a 0,49.

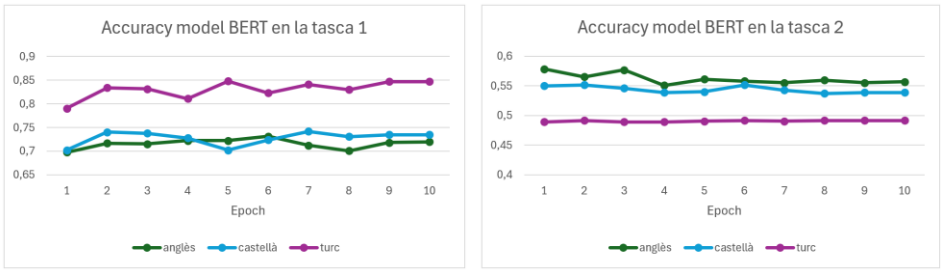


Figura 8: Resultats de la métrica accuracy per al model BERT emprant versions específiques per idioma

La Figura 9 correspon als resultats del *accuracy* obtinguts en el model BERT entrenat amb la versió multilingüe en les dues tasques propostes. Veiem el comportament dels quatre conjunts de dades: anglès, castellà, turc i tots els idiomes conjuntament. Els resultats obtinguts en els diferents idiomes és semblant a l'obtingut anteriorment per a les versions específiques. Remarcant els resultats obtinguts per al conjunt de dades que conté tots els idiomes en la tasca 1, ja que presenta un *accuracy* mitjà 0,77.

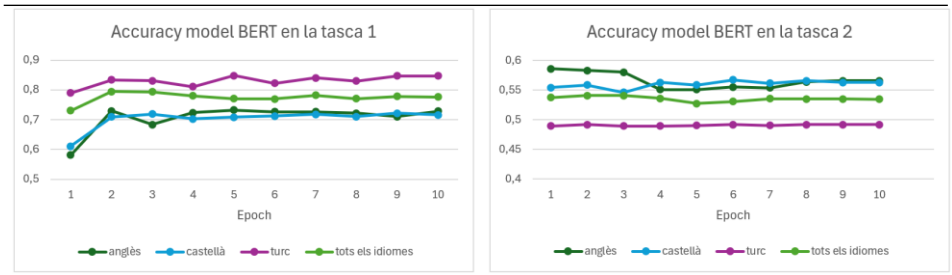


Figura 9: Resultats de la mètrica accuracy per al model BERT emprant la versió multilingüe

- **RoBERTa**: En el cas del model RoBERTa tal com veiem a la secció 4.6 Models de P, s'ha entrenat primerament amb versions específiques per l'anglès i el castellà i posteriorment amb una versió multilingüe per al conjunt de dades que conté tots els idiomes.

  A les següents gràfiques es reflecteixen els resultats obtinguts de la mètrica *loss* en cada *epoch*, avaluada a les dades d'entrenament i validació per a la tasca 1 (Figura 10: Resultats de la mètrica loss per a la tasca 1 en el model RoBERTa emprant versions específiques per idioma) i tasca 2 (Figura 11).

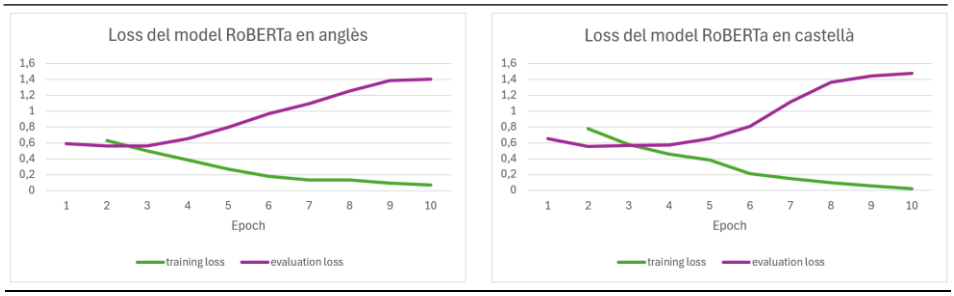


Figura 10: Resultats de la mètrica loss per a la tasca 1 en el model RoBERTa emprant versions específiques per idioma

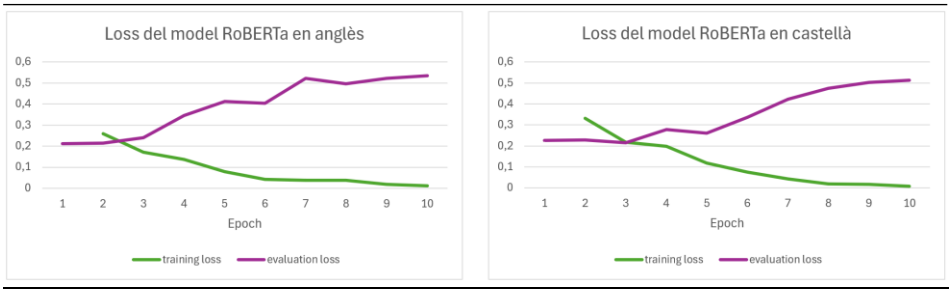


Figura 11: Resultats de la mètrica loss per a la tasca 2 en el model RoBERTa emprant versions específiques per idioma

La Figura 12 correspon als resultats del *accuracy* obtinguts en el model RoBERTa entrenat amb la versió per cadascun dels dos idiomes. On veiem que no és un comportament tan estable com el cas del model BERT. El *accuracy* mitjà en la classificació de textos sexistes és en els dos idiomes del 0,72 i en la categorització dels textos és del 0,55.

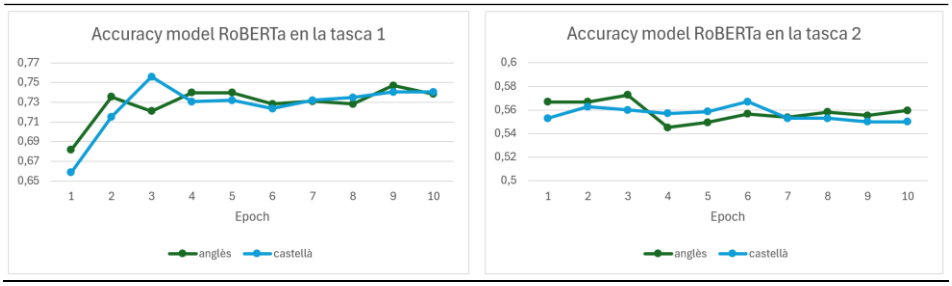


Figura 12: Resultats de la mètrica accuracy per al model RoBERTa emprant versions específiques per idioma

La Figura 13 correspon als resultats obtinguts per al model XLM RoBERTa entrenat amb tot el conjunt de dades que conté els tres idiomes. On es reflecteix un millor rendiment del model en la classificació dels textos (tasca 1) que en la categorització).

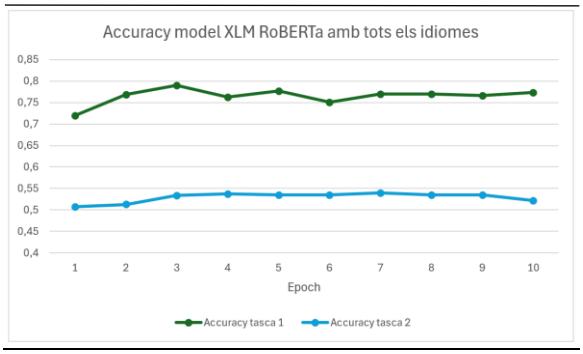


Figura 13: Resultats de la métrica accuracy per al model XLM RoBERTa

Per tant, podem concloure que els models funcionen bé amb les dades d'entrenament, però no s'ajusten correctament a les noves dades introduïdes com a validació del model. Per a millorar aquests resultats i evitar el overfitting del model, s'han portat a terme proves amb el número de epoch i ajustant el learning rate per a millorar la generalització. No obstant això, els resultats mostrats són els millors que s'han aconseguit.

Respecte a definir la millor mètrica escollida per a avaluar els models i com és el seu rendiment en les diferents versions és el *accuracy*. Tots els models testejats mostren un millor rendiment en la tasca de classificació dels textos en sexistes i no sexistes, respecte a la tasca de categorització dels textos en les diferents tipologies.

5.2 Limitacions de l'estudi

Al llarg del desenvolupament d'aquest treball s'han presentat diverses limitacions, afectant l'amplitud de l'estudi i els resultats obtinguts.

La limitació principal ha estat el temps d'execució dels models. Els models preentrenats emprats i desenvolupats al llarg del treball requereixen gran quantitat de recursos computacionals i temps per processar les dades. Encara tenir accés al HPC de la Universitat Pompeu Fabra i poder fer ús, no ha sigut prou suficient per manipular la càrrega de treball imposada pels models i els grans volums de dades. El qual ha sigut limitant a l'hora de poder dur a terme més proves de rendiment dels models i millorar els resultats obtinguts.

Una altra limitació significativa del treball va ser la manca de llibreries i models específics entrenats per a l'idioma turc. Mentre que l'anàlisi de textos en anglès i castellà compta amb una gamma més àmplia

d'eines i biblioteques, el turc no disposa de la mateixa varietat ni qualitat de recursos. Això ha obligat a testejar els models amb llibreries multilingües i no específiques amb aquest idioma.

En resum, encara que aquest treball s'ha pogut desenvolupar i acabar amb uns resultats rellevants, aquestes limitacions s'han de tenir en compte. Les investigacions futures podrien beneficiar-se de millores en la infraestructura computacional i el desenvolupament d'eines de PLN més robustes i específiques per a una varietat més àmplia d'idiomes específics.

Capítol 6

Interfície web interactiva

Hem implementat una interfície web per poder visualitzar d'una manera més pràctica el funcionament dels models. A més, aquesta web està creada amb el propòsit de crear un entorn més inclusiu i equitatiu mitjançant la identificació i eliminació del sexisme en totes les seves formes.

6.1 Arquitectura

La interfície web per a la classificació del sexisme s'ha desenvolupat utilitzant Streamlit, una eina de creació d'aplicacions web interactives a Python. L'aplicació permet als usuaris introduir text en un dels tres idiomes (anglès, castellà i turc) i mitjançant un dels models implementats al llarg del treball, classificar el contingut com a sexista o no sexista. En cas de ser classificat com a sexista, el text es categoritza segons les tipologies de sexisme explicades en l'apartat 3.2 Tipologies de sexisme.

Per a poder dur a terme la implementació de la interfície, hem guardat els resultats obtinguts a partir de l'entrenament de cada model. En funció de la selecció que realitza l'usuari amb l'idioma del text que vol analitzar i el model amb el qual desitja dur a terme l'anàlisi, carreguem el model pertinent. Aquesta selecció es realitza mitjançant dos menús desplegables.

L'usuari disposa d'una àrea de text on escriure el contingut que desitja analitzat. El text introduït per l'usuari es preprocessa de la mateixa manera que s'han processat totes les dades a l'hora d'entrenar els models, perquè sigui compatible amb el model.

A partir d'aquí, el model realitza una predicció de si el text es considera sexista o no. I en cas de ser un text sexista, el categoritza específicament. On els resultats es mostren a l'usuari a través de la interfície.

6.2 Funcionament de la interfície

És una plataforma senzilla d'utilitzar. Consta de dues pàgines:

- **Pàgina d'inici**: Aquesta pàgina conté una explicació del propòsit i com funciona l'aplicació.
- **Pàgina de classificació**: Aquesta pàgina permet a l'usuari introduir el text que desitja analitzar. I el sistema s'encarrega d'identificar qualsevol contingut sexista, proporcionant una anàlisi detallada i suggeriments per millorar la comunicació a la pàgina de resultats.

Amb aquesta web pots assegurar-te que el teu missatge sigui inclusiu i respectuós per a tothom.

En l'Apèndix 2: Exemple interfície web, trobem un exemple pràctic de com utilitzar la interfície i com analitza i classifica el text que introdueix un usuari.

Capítol 7

Conclusions

Aquest treball remarca la importància de detectar i categoritzar els textos sexistes a les xarxes socials perquè no es continuïn propagant actituds sexistes i discriminatòries.

Després de veure el rendiment de tots els models implementats en anglès, castellà i turc, podem concloure que els models de classificació simples han demostrat ser menys precisos, especialment al conjunt multilingüe, encara que els resultats obtinguts per al conjunt de dades en turc prou rellevants. Després de comparar tots els resultats, podem concloure que el tercer *pipeline* és el més bo per a la detecció i la categorització dels comentaris sexistes.

Pel que fa als models preentrenats d'aprenentatge profund, els resultats han mostrat ser més robustos, però han mostrat diferències significatives entre els resultats obtinguts en dur a terme les dues tasques proposades. En el cas de la tasca de classificar els textos, les mètriques analitzades mostren un bon rendiment del conjunt d'entrenament i fiabilitat dels models. Tot i això, en la tasca de categorització dels textos, els resultats obtinguts són similars al dels models simples.

Cal remarcar, que malgrat la manca de biblioteques i models específics entrenats per al turc i haver d'emprar eines multilingües en algun cas, els resultats obtinguts són rellevants.

En el cas de l'anglès i el castellà, el rendiment dels models al emprar llibreries específiques de l'idioma és inferior a emprar la versió multilingüe.

És crucial desenvolupar més eines i models específics per a una varietat més àmplia d'idiomes, incloent-hi aquells amb menys recursos disponibles com el turc. I implementar tècniques addicionals per prevenir *l'overfitting*, com la regularització i l'ús de conjunts de dades més àmplies i diverses, i així millorar la capacitat dels models per generalitzar.

En resum, aquest treball ha establert una base per a la detecció de llenguatge sexista utilitzant models simples de classificació i models avançats de PLN. Tot i que es van enfrontar diverses limitacions, els resultats obtinguts proporcionen una direcció per a futures investigacions i desenvolupaments en aquest camp.

Capítol 8

Projectes futurs

En aquest apartat es plantegen diverses línies de treball futur que poden abordar les limitacions actuals i obrir noves oportunitats de recerca en la detecció del sexisme a les xarxes socials:

- Realitzar una parametrització més extensa dels models utilitzats. Provar i ajustar els diferents paràmetres i configuracions per optimitzar el rendiment dels models i obtenir uns millors resultats. Una parametrització més detallada permetrà identificar les combinacions més efectives i millorar la capacitat del model per identificar i classificar comentaris sexistes amb més precisió.

- Com que els recursos existents per a l'idioma turc són més limitats que per a l'anglès, entrenar models com RoBERTA en una llibreria específica en turc permetrà més precisió i efectivitat la classificació.

- Realitzar una ampliació de la base de dades que contingui més idiomes. Per a poder augmentar la generalització i aplicabilitat dels models, és fonamental ampliar la base de dades, incloent-hi una varietat més àmplia de llengües. Aquesta expansió contribuirà a crear eines més globals i útils per a la detecció de sexisme a les xarxes socials.

- Desenvolupar sistemes de classificació automàtica i filtres basats en els models entrenats permetrà detectar i gestionar comentaris sexistes en temps real en les xarxes socials. La implementació de filtres pot ajudar les xarxes a tenir un entorn més segur i respectuós per a tots eles usuaris.

- Dur a terme l'anàlisi amb models de processament de llenguatge natural més avançats com per exemple LaMDA[23], WuDao 2.0 [24]o Megatron-Turing NLG[25].

- Millorar l'aplicació interactiva. Incloent noves funcionalitats, millores en la interfície d'usuari i optimitzar el rendiment per a utilitzar diversos idiomes.

Aquestes propostes futures poden superar les limitacions trobades en el desenvolupament del treball i expandir significativament l'abast i l'aplicabilitat de l'anàlisi del sexisme a les xarxes socials.

[23] LaMDA (Language Model for Dialogue Applications) és un model de llenguatge factual de Google AI, entrenat en un conjunt de dades massiu de text i codi. Pot generar text, traduir idiomes, escriure diferents tipus de contingut creatiu i respondre les vostres preguntes de manera informativa [80].

[24] WuDao 2.0 és un model de transformador dens només per a descodificadors de paràmetres d'1,75 bilions desenvolupat per l'Acadèmia d'Intel·ligència Artificial de Beijing (BAAI). És el model d'idioma més gran en xinès i s'ha demostrat que supera altres models lingüístics en una varietat de tasques de processament del llenguatge natural (PLN), com ara la traducció automàtica, la resposta a preguntes i el resum de text [81].

[25] Megatron-Turing NLG és un model de transformador dens només per a descodificadors de 530 mil milions de paràmetres desenvolupat per NVIDIA i Microsoft. És un dels models lingüístics més grans del món i s'ha demostrat que supera altres models lingüístics en una varietat de tasques de PLN, com ara la traducció automàtica, la resposta a preguntes i el resum de textos [82].

Referències bibliogràfiques

[1] Office of the High Commissioner for Human Rights. (s.f.). Women. United Nations Human Rights. Recuperat el 6 d'abril de 2024, de https://www.ohchr.org/en/women

[2] Velásquez Pérez, M. M. (2018). Estereotipos y roles de género en la educación inicial: Un estudio en instituciones educativas de Lima Metropolitana (Tesis de licenciatura, Pontificia Universidad Católica del Perú). Recuperat el 4 de maig de 2024, de https://tesis.pucp.edu.pe/repositorio/bitstream/handle/20.500.12404/12959/Vel%C3%A1squez_P%C3%A9rez_Estereotipos_roles_g%C3%A9nero1.pdf

[3] Concepto.de. (s.f.). Misoginia. Concepto.de. Recuperat el 12 de maig de 2024, de https://concepto.de/misoginia/

[4] United Nations. (s.f.). Gender equality. United Nations. Recuperat el 4 de junio de 2024, de https://www.un.org/es/global-issues/gender-equality

[5] Office of the High Commissioner for Human Rights. (s.f.). OHCHR homepage. United Nations Human Rights. Recuperat el 30 de març de 2024, de https://www.ohchr.org/en/ohchr_homepage

[6] United Nations. (s.f.). Gender equality. United Nations. Recuperat el 12 de febrer de 2024, de https://www.un.org/es/global-issues/gender-equality

[7] Rodríguez-Sánchez, F., Carrillo-De-Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., & Donoso, T. (2021, 6 septiembre). Overview of EXIST 2021: sEXism Identification in Social neTworks. Rodríguez-Sánchez | Procesamiento del Lenguaje Natural. https://rua.ua.es/dspace/bitstream/10045/117491/1/PLN_67_17.pdf

[8] EXIST 2022. (2022). EXIST: sEXism Identification in Social neTworks Second Shared Task at IberLEF 2022. Recuperat el 3 de febrer de 2022, de http://nlp.uned.es/exist2022/

[9] UNED. (s.f.). eXist 2023. UNED. Recuperat el 1 juny de 2024, de http://nlp.uned.es/exist2023/

[10] UNED. (s.f.). eXist 2024. UNED. Recuperat el 1 juny de 2024, de http://nlp.uned.es/exist2024/

[11] World Health Organization. (s.f.). Violence against women. World Health Organization. Recuperat el 3 de març de 2024, de https://www.who.int/es/news-room/fact-sheets/detail/violence-against-women

[12] Marcha Mundial de las Mujeres. (s.f.). Objetivos de la Marcha Mundial. Marcha Mundial de las Mujeres. Recuperat el 1 de maig de 2024, de https://marchemondiale.org/index.php/quienes-somos/objetivos-de-la-marcha-mundial/?lang=es

[13] #MeToo and #BalanceTonPorc: Translating Feminism. (2023, 6 febrero). The American University Of Paris. Recuperat el 13 de maig de 2024, de https://www.aup.edu/news-events/news/2018-03-09/metoo-and-balancetonporc-translating-feminism

[14] Manzano-Zambruno, L. (s. f.). ¿Es el #MeToo un movimiento? Una revisión sobre el concepto «movimiento social» y su relación con las redes sociales. *Comunicación y Pensamiento*. Recuperat el 5 de maig de 2024, de https://comunicacionypensamiento.org/ediciones/2019/ponencia/es-el-metoo-un-movimiento-una-revision-sobre-el-concepto-movimiento-social-y-su-relacion-con-las-redes-sociales/index.html

[15] Pérez, P. A. (2018, 13 octubre). Todos los «weinstein» sin desenmascarar: el #BalanceTonPorc francés cumple un año. publico.es. Recuperat el 21 de abril de 2024, de https://www.publico.es/sociedad/violencia-machista-weinstein-desenmascarar-balancetonporc-frances-cumple-ano.html

[16] Bela-Lobedde, D., & Bela-Lobedde, D. (2020, 16 noviembre). #SayHerName también en español. Desenredando. Recuperat el 11 de maig de 2024, de https://blogs.publico.es/desenredando/2020/11/17/sayhername-tambien-en-espanol/?doing_wp_cron=1715351665.8028919696807861328125

[17] Ministerio de Igualdad. (s.f.). Definición de violencia de género. Portal de la Violencia de Género. Recuperat el 28 de març de 2024, de https://violenciagenero.igualdad.gob.es/definicion/

[18] Equal Justice Initiative. (s.f.). Racial justice. Equal Justice Initiative. Recuperat el 11 de maig de 2024, de https://eji.org/racial-justice/

[19] Zou, J., et al. (2017). Tweeting offensive language: A fine-grained analysis and classification using distant supervision. arXiv preprint arXiv:1702.07179.

[20] Davidson, T., et al. (2018). Offensive language detection in online social media. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 61-69.

[21] Yıldırım, G., et al. (2019). Convolutional neural networks for sentiment analysis of Turkish tweets. In Proceedings of the International Conference on Artificial Intelligence and Data Science, pp. 308-317.

[22] Sanches-Piqueras, J., et al. (2019). Detección de mensajes sexistas en Twitter en español. In Proceedings of the XLVII Congreso de la Sociedad Española de Lingüística, pp. 775-782.

[23] Díaz-Galiano, J., et al. (2020). Clasificación de comentarios ofensivos en foros de noticias en español. In Proceedings of the XXVIII Congreso Español de Inteligencia Artificial, pp. 337-344.

[24] Huerta, E., et al. (2021). Identification of sexist tweets in Spanish using deep learning with recurrent neural networks. In Proceedings of the 2021 Conference on Artificial Intelligence Research & Development, pp. 161-168.

[25] Altın, L. S. M., & Saggion, H. (2024b, mayo 1). *A Novel Corpus for Automated Sexism Identification on Social Media*. ACL Anthology. https://aclanthology.org/2024.sigul-1.2/

[26] Tocoglu, M. A., Ozturkmenoglu, O., & Alpkocak, A. (2019). Emotion analysis from Turkish tweets using deep neural networks. *IEEE Access*, *7*, 183061-183069. https://doi.org/10.1109/access.2019.2960113

[27] Miran, A. Z., & Yahia, H. S. (2023). Hate Speech Detection in Social Media (Twitter) Using Neural Network. *Journal of Mobile Multimedia*, February 2023.

[28] Sexo. (s. f.). Inmujeres. Recuperat el 3 de febrer de 2024, https://campusgenero.inmujeres.gob.mx/glosario/terminos/sexo

[29] Real Academia Española. (s.f.). Género. Diccionario de la lengua española. Recuperat el 9 de maig de 2024, de https://dle.rae.es/g%C3%A9nero

[30] Mundo Psicólogos. (s.f.). Identidad sexual. Mundo Psicólogos. Recuperat el 9 de maig de 2024, de https://www.mundopsicologos.com/diccionario-psicologico/identidad-sexual

[31] Definición de sexismo. origen, historia, manifestaciones y conceptos. (s. f.-b). https://definicion.com/sexismo/

[32] Inmujeres. https://campusgenero.inmujeres.gob.mx/glosario/terminos/sexismo

[33] Rigol, M. (2018, 17 octubre). *Desigualdad de género - Fundació Surt*. Fundació Surt. https://www.surt.org/es/desigualdad-de-genero/

[34] Aranda Martínez, E. (2023). GUIA PRÀCTICA DE LLENGUATGE INCLUSIU i NO SEXISTA. En Institut Andorrà de Les Dones (ISBN 978-99920-3-393-7). Recuperat 6 de marzo de 2024, de https://irp.cdn-website.com/cb103d82/files/uploaded/Guia%20practica%20de%20llenguatge%20inclusiu.pdf

[35] Sexisme. Detecta'l. Posa-li nom. Atura'l. (s. f.). Human Rights Channel. https://human-rights-channel.coe.int/stop-sexism-ca.html

[36] La misoginia y el sexismo se agudizan en el mundo, advierte experta. (2021, 22 octubre). Noticias ONU. https://news.un.org/es/story/2021/10/1498542

[37] Instituto Nacional de las Mujeres. (s.f.). Patriarcado. Recuperat el 17 d'abril de 2024, de https://campusgenero.inmujeres.gob.mx/glosario/terminos/patriarcado

[38] Alfonso Chaves-Montero, Alfonso (2020). ¿Cuáles son los roles y estereotipos de género sexistas de la sociedad? Recuperat el 29 de mar´de 2024, de https://2020.nodos.org/ponencia/cuales-son-los-roles-y-estereotipos-de-genero-sexistas-de-la-sociedad

[39] Saavedra, N., & Casal, L. (s. f.). Género, sexismo y variables culturales. En Universidad de la Coruña.

[40] Rengel, C. (2023, 25 noviembre). Los países más machistas y menos avanzados en igualdad de género en la UE. ElHuffPost. https://www.huffingtonpost.es/politica/los-paises-mas-machistas-avanzados-igualdad-genero-ue.html

[41] Economía feminista | Exploring Economics. (s. f.). https://www.exploring-economics.org/es/orientacion/feminist-economics/

[42] BOE-A-2022-11589 Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación. (s. f.). https://www.boe.es/buscar/act.php?id=BOE-A-2022-11589

[43] La Violencia Contra las Mujeres, C. N. P. P. y. E. (s. f.). Entrevista: Sexismo en los medios de comunicación: qué es y cómo a. . . gob.mx. https://www.gob.mx/conavim/es/articulos/entrevista-sexismo-en-los-medios-de-comunicacion-que-es-y-como-afecta-a-las-mujeres?idiom=es

[44] Umaximo. (2020). Educación sexista: ¿Qué es? Umaximo. Recuperat el 9 de maig de 2024, de https://www.umaximo.com/post/educacion-sexista-que-es

[45] Educación 3.0. (2023). Sexismo en la escuela. Educación 3.0. Recuperat el 29 de febrer de 2024, de https://www.educaciontrespuntocero.com/noticias/sexismo-en-la-escuela/

[46] Mundiario. (2019, octubre 3). Investigación encuentra que el sexismo puede dañar la salud física y mental de las mujeres. Mundiario. Recuperat el 1 de maig de 2024, de https://www.mundiario.com/articulo/mundilife/investigacion-encuentra-sexismo-puede-danar-salud-fisica-mental-mujeres/20191003174708165526.html

[47] Sensacionweb. (s.f.). Stop word: Qué es, definición, significado y ejemplos. Sensacionweb. Recuperat el 6 de maig de 2024, de https://sensacionweb.com/diccionario/stop-word-que-es-definicion-significado-y-ejemplos/

[48] DataScientest. (s.f.). NLP: Introducción. DataScientest. Recuperat el 21 defebrer de 2024, de https://datascientest.com/es/nlp-introduccion

[49] Hugging Face community. (n.d.). Summary of the tokenizers. Recuperat el 21 de abril de 2024, de from https://huggingface.co/docs/transformers/tokenizer_summary

[50] DataScientest. (s.f.). NLP: Introducción. DataScientest. Recuperat el 4 de maig de 2024, de https://datascientest.com/es/nlp-introduccion

[51] Hugging Face. (s.f.). Glossary. Hugging Face. Recuperat el 11 de febrer de 2024, de https://huggingface.co/transformers/v3.1.0/glossary.html

[52] *Google Colab*. (s. f.-b). https://research.google.com/colaboratory/faq.html?authuser=0&hl=es

[53] Universitat Pompeu Fabra. (s.f.). Home - HPC Guide. Guies BibTIC. Recuperat el 8 de març de 2024, de https://guiesbibtic.upf.edu/recerca/hpc/home

[54] Blanca, A. C., & Blanca, A. C. (2023, 23 noviembre). Explicación del Teorema de Bayes: fórmula y ejemplos. Academia Carta Blanca. https://academiacartablanca.es/blog/teorema-de-bayes-probabilidad/

[55] Mathematics. (s.f.). Adapting Hidden Naive Bayes for Text Classification. *mdpi.com*. Recuperat el 11 de abril de 2024, de https://www.mdpi.com/journal/mathematics

[56] Patel, H. (2019, 18 diciembre). Text classification using K Nearest Neighbors (KNN). OpenGenus IQ: Learn Algorithms, DL, System Design. https://iq.opengenus.org/text-classification-using-k-nearest-neighbors/

[57] Huang, A., Xu, R., Chen, Y., & Guo, M. (2023). Research on multi-label user classification of social media based on ML-KNN algorithm. *Technological Forecasting & Social Change/Technological Forecasting And Social Change*, *188*, 122271. https://doi.org/10.1016/j.techfore.2022.122271

[58] Luo, X. (2017). A New Text Classifier Based on Random Forests. Atlantis Press. https://doi.org/10.2991/meita-16.2017.60

[59] Kumar, P., & Wahid, A. (2021). Social Media Analysis for Sentiment Classification Using Gradient Boosting Machines. En Algorithms for intelligent systems (pp. 923-934). https://doi.org/10.1007/978-981-16-3246-4_70

[60] Python, R. (2023, 26 junio). Logistic Regression in Python. https://realpython.com/logistic-regression-python/

[61] Sharma, P. (2021, 28 abril). Python Sklearn Logistic Regression Tutorial with Example - MLK - Machine Learning Knowledge. MLK - Machine Learning Knowledge. https://machinelearningknowledge.ai/python-sklearn-logistic-regression-tutorial-with-example/

[62] ¿Qué es el procesamiento del lenguaje natural (PLN)? | IBM. (s. f.). https://www.ibm.com/mx-es/topics/natural-language-processing

[63] Effrosynidis, Dimitris. (2024, 20 febrero). The Transformer Architecture From a Top View. Towards AI. https://towardsai.net/p/machine-learning/the-transformer-architecture-from-a-top-view

[64] Ferrer, J. (2024). How Transformers Work: A detailed exploration of transformer architecture. En Datacamp. https://www.datacamp.com/tutorial/how-transformers-work

[65] De Luis, A. (2023, 11 mayo). Modelo BERT para hacer trading por Quantpedia. Método Trading. https://metodotrading.com/modelo-bert/

[66] BERT 101 - state of the art NLP model explained. (s. f.). https://huggingface.co/blog/bert-101

[67] Hugging Face. (s.f.). Google BERT base uncased. Recuperat el 26 de maig de 2024, de https://huggingface.co/google-bert/bert-base-uncased

[68] Hugging Face. (s.f.). Modelo dccuchile/bert-base-spanish-wwm-uncased. Recuperat el 12 de febrer de 2024, de https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased

[69] DBMDZ. (s.f.). Modelo dbmdz/bert-base-turkish-cased. Recuperat el 30 de gener de 2024, de https://huggingface.co/dbmdz/bert-base-turkish-cased

[70] google-bert/bert-base-multilingual-cased · Hugging Face. (2001, 11 marzo). Recuperat el 30 de gener de 2024, de https://huggingface.co/google-bert/bert-base-multilingual-cased

[71] Chandan Durgia. (2021). Exploring BERT variants (Part 1): ALBERT, RoBERTa, ELECTRA. In Towards Data Science . Medium. Recuperat el 2 de maig de 2024, de https://towardsdatascience.com/exploring-bert-variants-albert-roberta-electra-642dfe51bc23

[72] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. Recuperat el 2 de maig de 2024, de https://arxiv.org/abs/1907.11692

[73] RoBERTa. (s. f.). https://huggingface.co/docs/transformers/model_doc/roberta

[74] Facebook AI. (s.f.). Modelo roberta-base. Hugging Face. Recuperat el 20 de març de 2024, de https://huggingface.co/FacebookAI/roberta-base

[75] PlanTL-GOB-ES. (s.f.). Modelo roberta-base-bne. Hugging Face. Recuperat el 17 de març de 2024, de https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne

[76] Hugging Face. (s.f.). Modelo xlm-roberta-base. Recuperat el 5 de juny de 2024, de https://huggingface.co/xlm-roberta-base

[77] What Is a Machine Learning Pipeline? | IBM. (s. f.). https://www.ibm.com/topics/machine-learning-pipeline

[78] Marcello Politi. (2022). Fine-Tuning for Domain Adaptation in NLP. Towards Data Science Towards Data Science Your Home for Data Science. A Medium Publication Sharing Concepts, Ideas and Codes. Towards Data Science Followed by 663905 People Follow. Recuperat el 22 de març de 2023, de https://towardsdatascience.com/fine-tuning-for-domain-adaptation-in-nlp-c47def356fd6

[79] B, H. N. (2021, 12 diciembre). Confusion matrix, accuracy, precision, recall, F1 score. Medium. https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd

[80] Google. (2021, mayo 19). Introducing LaMDA: A language model for dialogue. Blog de Google. Recuperat el 4 de maig de 2024, de https://blog.google/technology/ai/lamda/

[81] Baidu Research. (s.f.). Título del artículo. Baidu Research. Recuperat el 18 de març de 2024, de http://research.baidu.com/Blog/index-view?id=178

[82] NVIDIA Developer. (2023, marzo 2). Using DeepSpeed and Megatron to train Megatron Turing NLG-530B, the world's largest and most powerful generative language model. NVIDIA Developer Blog. Recuperat de https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

[83] Guies BiBTIC: HPC High Performance Computing: 8.2. JuPyter Notebook. (s. f.). https://guiesbibtic.upf.edu/recerca/hpc/jupyter-notebook

Apèndix 1: Resultats models de classificació

En aquest apèndix trobem els resultats obtinguts en l'entrenament dels models de classificació en les diferents classificacions. En primer lloc, trobem les taules de resultats obtingudes en la tasca de classificar els textos en sexistes o no sexistes (tasca 1) i, en segon lloc, trobem les taules de resultats obtingudes en la categorització dels textos sexistes.

Naive Bayes	Idioma	Mètriques			
		Accuracy	Precision	Recall	F1 score
Pipeline 1	Anglés	0,63	0,66	0,64	0,63
	Català	0,67	0,68	0,67	0,67
	Turc	0,59	0,6	0,6	0,57
	Tots els idiomes	0,49	0,46	0,49	0,44
Pipeline 2	Anglés	0,67	0,67	0,67	0,67
	Català	0,69	0,7	0,7	0,7
	Turc	0,78	0,79	0,79	0,78
	Tots els idiomes	0,67	0,69	0,68	0,68
Pipeline 3	Anglés	0,67	0,68	0,67	0,67
	Català	0,69	0,69	0,69	0,69
	Turc	0,79	0,81	0,79	0,78
	Tots els idiomes	0,71	0,71	0,71	0,71

Taula 6: Resultats del model Naive Bayes per a la tasca 1

Naive Bayes	Idioma	Mètriques			
		Accuracy	Precision	Recall	F1 score
Pipeline 1	Anglés	0,56	0,56	0,56	0,52
	Català	0,58	0,64	0,58	0,53
	Turc	0,61	0,6	0,61	0,56
	Tots els idiomes	0,54	0,53	0,54	0,44
Pipeline 2	Anglés	0,52	0,54	0,53	0,42
	Català	0,54	0,6	0,54	0,44
	Turc	0,66	0,64	0,66	0,61
	Tots els idiomes	0,57	0,59	0,57	0,52
Pipeline 3	Anglés	0,48	0,49	0,48	0,32
	Català	0,49	0,57	0,49	0,33
	Turc	0,59	0,68	0,59	0,48
	Tots els idiomes	0,52	0,66	0,52	0,39

Taula 7: Resultats del model Naive Bayes per a la tasca 2

KNN	Idioma	Mètriques			
		Accuracy	Precision	Recall	F1 score
Pipeline 1	Anglés	0,64	0,65	0,64	0,64
	Català	0,62	0,72	0,64	0,62
	Turc	0,78	0,79	0,78	0,78
	Tots els idiomes	0,62	0,7	0,62	0,59
Pipeline 2	Anglés	0,61	0,61	0,61	0,61
	Català	0,58	0,63	0,61	0,6
	Turc	0,57	0,72	0,57	0,51
	Tots els idiomes	0,56	0,67	0,56	0,5
Pipeline 3	Anglés	0,64	0,63	0,63	0,63
	Català	0,63	0,63	0,63	0,63
	Turc	0,52	0,75	0,52	0,42
	Tots els idiomes	0,51	0,74	0,5	0,37

Taula 8: Resultats del model KNN per a la tasca 1

KNN	Idioma	Mètriques			
		Accuracy	Precision	Recall	F1 score
Pipeline 1	Anglés	0,55	0,54	0,55	0,52
	Català	0,54	0,57	0,55	0,51
	Turc	0,61	0,6	0,61	0,58
	Tots els idiomes	0,19	0,35	0,19	0,16
Pipeline 2	Anglés	0,45	0,5	0,45	0,42
	Català	0,52	0,59	0,52	0,48
	Turc	0,54	0,58	0,54	0,45
	Tots els idiomes	0,52	0,56	0,52	0,43
Pipeline 3	Anglés	0,51	0,49	0,51	0,49
	Català	0,52	0,51	0,52	0,49
	Turc	0,52	0,61	0,52	0,4
	Tots els idiomes	0,5	0,67	0,5	0,36

Taula 9: Resultats del model KNN per a la tasca 2

Random forest	Idioma	Mètriques			
		Accuracy	Precision	Recall	F1 score
Pipeline 1	Anglés	0,7	0,73	0,68	0,7
	Català	0,69	0,78	0,58	0,67
	Turc	0,81	0,83	0,82	0,83
	Tots els idiomes	0,71	0,76	0,67	0,71
Pipeline 2	Anglés	0,71	0,7	0,7	0,7
	Català	0,69	0,71	0,7	0,69
	Turc	0,84	0,84	0,84	0,84
	Tots els idiomes	0,75	0,76	0,76	0,76
Pipeline 3	Anglés	0,7	0,72	0,72	0,72
	Català	0,69	0,72	0,71	0,7
	Turc	0,83	0,84	0,84	0,84
	Tots els idiomes	0,76	0,76	0,76	0,76

Taula 10: Resultats del model Random forest per a la tasca 1

Random Forest	Idioma	Mètriques			
		Accuracy	Precision	Recall	F1 score
Pipeline 1	Anglés	0,56	0,55	0,56	0,54
	Català	0,61	0,61	0,61	0,58
	Turc	0,68	0,66	0,68	0,66
	Tots els idiomes	0,61	0,59	0,61	0,57
Pipeline 2	Anglés	0,56	0,6	0,56	0,5
	Català	0,58	0,65	0,58	0,51
	Turc	0,71	0,69	0,71	0,68
	Tots els idiomes	0,64	0,63	0,64	0,61
Pipeline 3	Anglés	0,56	0,59	0,56	0,49
	Català	0,59	0,67	0,59	0,4
	Turc	0,71	0,69	0,71	0,68
	Tots els idiomes	0,64	0,64	0,64	0,61

Taula 11: Resultats del model Random Forest per a la tasca 2

Gradient boosting	Idioma	Mètriques			
		Accuracy	Precision	Recall	F1 score
Pipeline 1	Anglés	0,67	0,76	0,56	0,64
	Català	0,7	0,74	0,64	0,69
	Turc	0,81	0,83	0,81	0,82
	Tots els idiomes	0,69	0,79	0,59	0,67
Pipeline 2	Anglés	0,67	0,7	0,68	0,68
	Català	0,69	0,7	0,7	0,7
	Turc	0,81	0,82	0,82	0,82
	Tots els idiomes	0,71	0,73	0,71	0,71
Pipeline 3	Anglés	0,69	0,71	0,71	0,71
	Català	0,71	0,71	0,7	0,69
	Turc	0,82	0,82	0,82	0,82
	Tots els idiomes	0,72	0,73	0,72	0,72

Taula 12: Resultats del model Gadient boosting per a la tasca 1

Gradient Boosting	Idioma	Mètriques			
		Accuracy	Precision	Recall	F1 score
Pipeline 1	Anglés	0,57	0,57	0,58	0,54
	Català	0,63	0,65	0,63	0,6
	Turc	0,58	0,66	0,68	0,65
	Tots els idiomes	0,59	0,62	0,59	0,53
Pipeline 2	Anglés	0,57	0,57	0,57	0,53
	Català	0,63	0,66	0,63	0,58
	Turc	0,58	0,66	0,68	0,65
	Tots els idiomes	0,62	0,63	0,62	0,57
Pipeline 3	Anglés	0,58	0,57	0,58	0,54
	Català	0,63	0,64	0,63	0,59
	Turc	0,68	0,66	0,58	0,66
	Tots els idiomes	0,62	0,64	0,62	0,57

Taula 13: Resultats del model Gadient boosting per a la tasca 2

Regressió logística	Idioma	Mètriques			
		Accuracy	Precision	Recall	F1 score
Pipeline 1	Anglés	0,7	0,75	0,62	0,68
	Català	0,7	0,78	0,57	0,66
	Turc	0,79	0,85	0,76	0,8
	Tots els idiomes	0,71	0,79	0,59	0,68
Pipeline 2	Anglés	0,72	0,76	0,68	0,71
	Català	0,71	0,76	0,63	0,69
	Turc	0,82	0,85	0,83	0,84
	Tots els idiomes	0,75	0,79	0,72	0,76
Pipeline 3	Anglés	0,71	0,75	0,66	0,68
	Català	0,71	0,78	0,63	0,66
	Turc	0,84	0,82	0,9	0,8
	Tots els idiomes	0,76	0,72	0,76	0,68

Taula 14: Resultats del model Regressió logística per a la tasca 1

Regressió logística	Idioma	Mètriques			
		Accuracy	Precision	Recall	F1 score
Pipeline 1	Anglés	0,58	0,57	0,58	0,54
	Català	0,59	0,62	0,6	0,55
	Turc	0,61	0,6	0,61	0,57
	Tots els idiomes	0,55	0,56	0,55	0,47
Pipeline 2	Anglés	0,57	0,55	0,57	0,55
	Català	0,61	0,6	0,61	0,57
	Turc	0,68	0,66	0,68	0,66
	Tots els idiomes	0,62	0,61	0,63	0,6
Pipeline 3	Anglés	0,56	0,58	0,56	0,5
	Català	0,58	0,65	0,59	0,52
	Turc	0,68	0,65	0,68	0,64
	Tots els idiomes	0,62	0,64	0,62	0,58

Taula 15: Resultats del model Regressió logística per a la tasca 2

Apèndix 2: Exemple interfície web

En aquest apèndix veiem les captures realitzades a la interfície amb un exemple pràctic de classificació.

Figura 14: Pàgina d'inici

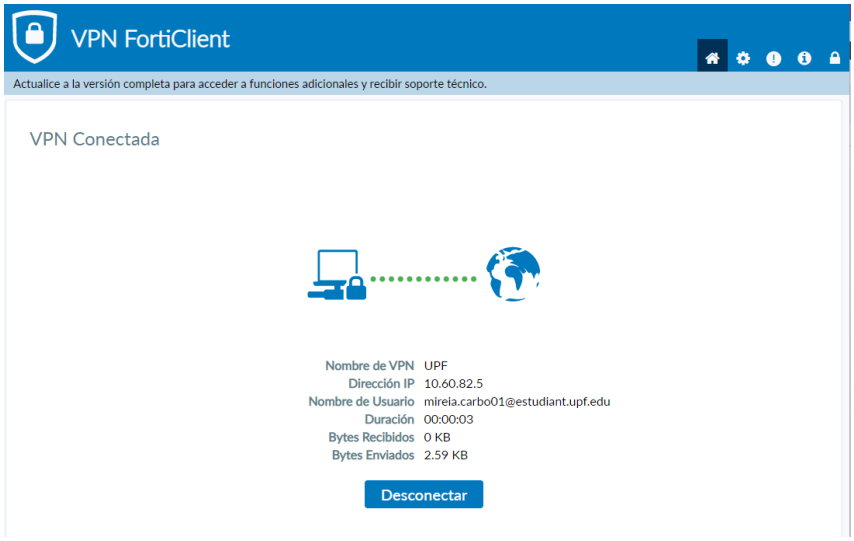
Figura 15: Pàgina de classificació

Figura 16: Pàgina de classificació amb resultats del comentari

Apèndix 3: Connexió al HPC

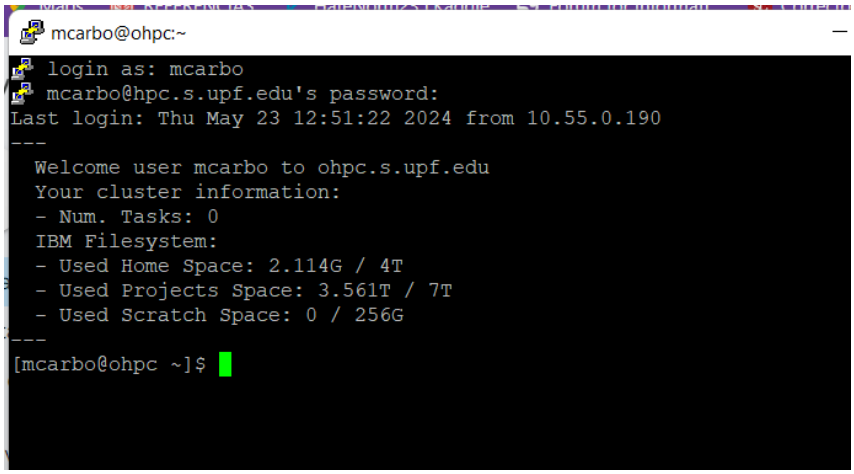
En aquest apèndix trobem els passos a seguir per a realitzar la connexió al HPC de la universitat [83].

1- Realitzar la connexió a la VPN de la UPF amb FortiClient



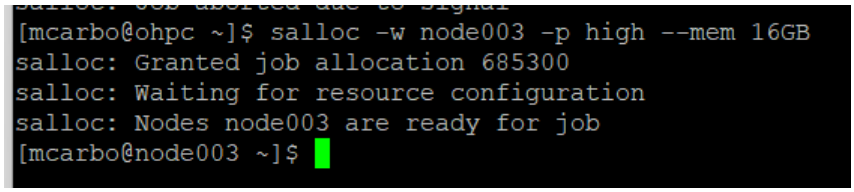
Il·lustració 1: Connexió FortiClient

2- Obrir putty i connectar-se al servidor hpc.s.upf.edu



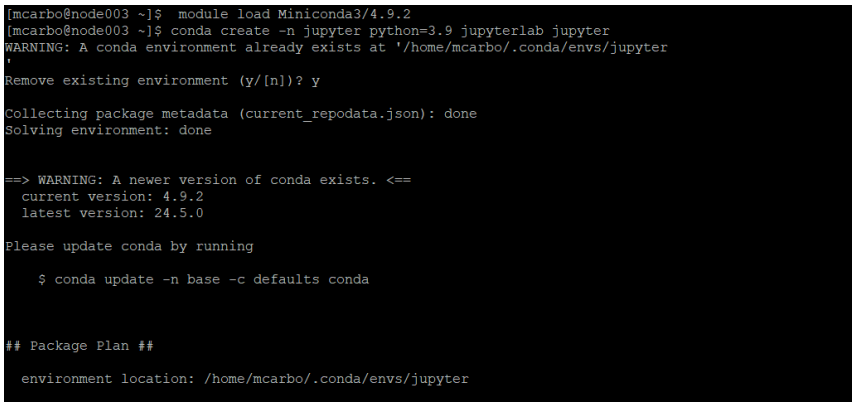
Il·lustració 2: Realitzar la connexió al servidor

3- Sol·licita una sessió interactiva en el node node003 per exemple, en la partició high, amb una assignació de 16 GB de memòria. Un cop que els recursos estiguin disponibles es connecta al node.



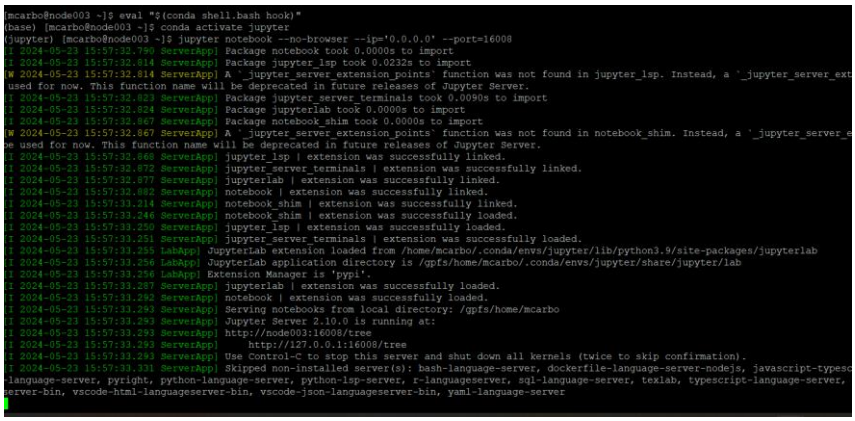
Il·lustració 3: connexió al node

4- Carregar Miniconda3



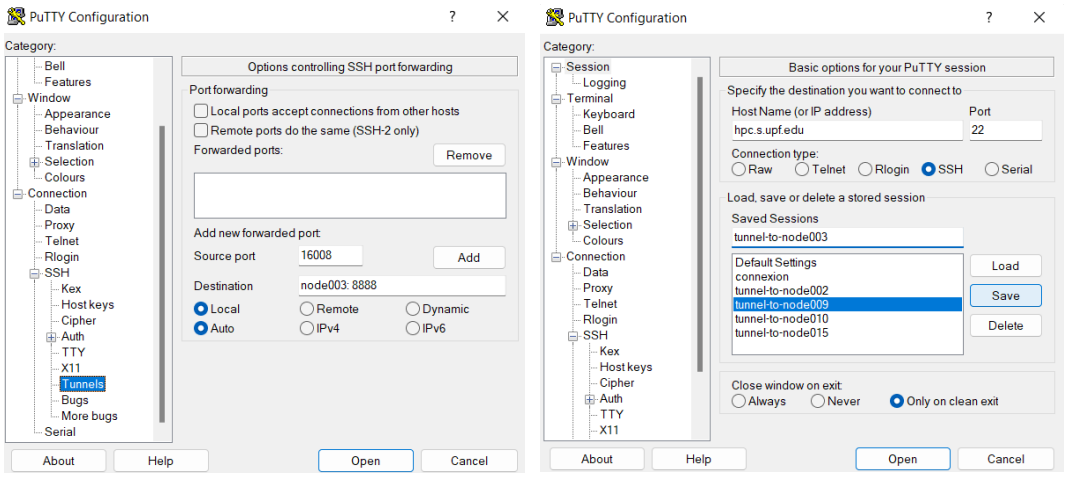
Il·lustració 4: Activar Miniconda

5- Carregar el bash shell i activar l'entorn



Il·lustració 5: Activació de l'entorn

6- Obrir una altra terminal per realitzar la connexió al nostre port físic i executem (estableix una connexió SSH al servidor hpc.s.upf.edu amb el nom d'usuari nsurname, i configura un túnel SSH que redirigeix el trànsit del port local 16008 a la nostra màquina al port 8888 a través del servidor hpc.s.upf.edu.)



Il·lustració 6: Tunnel del port a la màquina

7- Obrim el buscador i executem: `http://localhost:16008/`

Apèndix 4: Abreviatures

Al llarg del treball hem emprat diferents abreviatures, les definim en la següent taula.

Abreviatura	Nom en anglès	Nom en català
AAPF	Association for the Advancement of Black Women and Girls	Associació per a l'Avanç de les Dones i Nenes Negres
ANN	Artificial Neural Networks	Xarxes Neuronals Artificials
API	Application Programming Interface	Interfície de Programació d'Aplicacions
BERT	Bidirectional Encoder Representations from Transformers	Representacions Bidireccionals d'Encoders de Transformers
CISPS	Centre for Social Policy Studies and Research	Centre d'Estudis de Política Social i Investigació
CLEF	Cross-Language Evaluation Forum	Fòrum d'Avaluació Interlingüístic
CNN	Convolutional Neural Networks	Xarxes Neuronals Convolucionals
EXIST	sEXism Identification in Social networks	Identificació del Sexisme en Xarxes Socials
FN	False Negatives	Falsos Negatius
FP	False Positives	Falsos Positius
HCP	High Performance Computing	Computació d'Altes Prestacions
IDF	Inverse Document Frequency	Freqüència Inversa de Document
IEG-UC3M	Institute for Gender Studies, University Carlos III of Madrid	l'Institut d'Estudis de Gènere de la Universitat Carlos III de Madrid
JSON	JavaScript Object Notation	Notació d'Objectes JavaScript

KNN	K-Nearest Neighbors	K-Nearest Neighbors
LSTM	Long Short-Term Memory	Memòria a Curt i Llarg Termini
MNB	Multinomial Naive Bayes	Naive Bayes Multinomial
PLN	Natural Language Processing	Processament de Llenguatge Natural
RNN	Recurrent Neural Networks	Xarxes Neuronals Recurrents
RoBERTa	Robustly optimized BERT approach	Enfocament BERT optimitzat de manera robusta
SVM	Support Vector Machines	Màquines de Suport Vectorial
TF	Term Frequency	Freqüència de Terme
TN	True Negatives	Negatius Reals
TP	True Positives	Positius Reals
TURTED	Turkish Twitter Emotion Dataset	Conjunt de Dades d'Emocions a Twitter en Turc
VPN	Virtual Private Network	Xarxa Privada Virtual
XLM-RoBERTa	Cross-lingual Language Model RoBERTa	Model de Llenguatge Multilingüe RoBERTa

Taula 16: Abreviatures del treball