



UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH  
Facultat d'Informàtica de Barcelona



FACULTAT D'INFORMÀTICA DE BARCELONA  
GRAU EN ENGINYERIA INFORMÀTICA  
ESPECIALITAT DE TECNOLOGIES DE LA INFORMACIÓ

# **Eina de processat del llenguatge contra el ciberbullying**

TREBALL DE FINAL DE GRAU  
MEMÒRIA

**Natalia Lasheras Torrella**

Director: Manel Medina Llinàs  
Tutor GEP: Joaquim Deulofeu Aymar

Dilluns, 21 Juny de 2021

## Resum

L'increment de casos de ciberbullying en la actualitat i especialment en menors és un fenomen preocupant que ve donat per l'ús constant i en augment de les xarxes socials així com l'anonimat que aquestes s'ofereixen. Aquest projecte busca explorar la possibilitat de detecció de ciberbullying en missatges escrits en català, no per paraules claus sinó entenent la intencionalitat de la oració escrita, per a oferir una alternativa més efectiva que aquesta i més innovadora.

Per a fer-ho es busca una solució orientada en el camp de deep learning de manera que es generi un model que mitjançant tècniques del processat de llenguatge natural pugui determinar la probabilitat de que el missatge analitzat tingui intencionalitat d'assetjar online. Així doncs, es genera el nostre propi data set així com el nostre propi model el qual es contrasta amb altres classificadors existents per a poder realitzar un anàlisi i generar un model que doni solució a la qüestió amb el major percentatge de predicció possible.

## Resumen

El incremento de casos de ciberbullying en la actualidad y especialmente en menores es un fenómeno preocupante que viene dado del uso constante y en aumento de las redes sociales así como el anonimato que estas ofrecen. Este proyecto busca explorar la posibilidad de detección de ciberbullying en mensajes escritos en catalán, no por palabras clave sino entendiendo la intencionalidad de la oración escrita, para ofrecer una alternativa más efectiva e innovadora que esta.

Para hacerlo se busca una solución orientada en el campo de deep learning de modo que se genere un modelo que mediante técnicas del procesamiento del lenguaje natural se pueda determinar la probabilidad de que el mensaje analizado tenga la intención de acosar. Así pues, se genera nuestro propio dataset y nuestro propio modelo que será contrastado con otros clasificadores ya existentes para poder realizar un análisis y generar un modelo que dé solución a la cuestión con el mejor porcentaje de predicción posible.

## Abstract

The increase of cases of ciberbullying in the current climate and specially in under age kids, is a worrisome phenomenon brought on by the constant and rising use of social networks and the anonymity those provide to perpetrators. This project searches to explore the possibility of detection of ciberbullying situations in messages written in Catalan, not by the identification of key words but by understanding the intention behind the text written, in order to offer a better and more innovative alternative to the solutions explored nowadays.

In order to do so we search for a solution focused in the field of deep learning in order to generate a model that by the use of natural processing of the language techniques allows us to determine the probability of a message being of ill-intentioned and therefore creating a new tool to battle this ever present situation. We do so by creating our own dataset as well as our own predictive model which will be contrasted with different pre-existent classifiers will be analyzed so we can create a model that solves this problematic with the highest success rate possible.

## Taula de contingut

|       |   |    |
|-------|---|----|
| 1.    | Introducció i contextualització.....      | 7  |
| 1.1   | Contextualització.....                    | 7  |
| 1.2   | Definició de conceptes .....              | 8  |
| 1.3   | Identificació del problema .....          | 9  |
| 1.4   | Actors Implicats.....                     | 10 |
| 2.    | Justificació .....                        | 10 |
| 2.1   | Situació actual .....                     | 10 |
| 2.2   | Solucions ja existents .....              | 11 |
| 2.3   | Solució decidida.....                     | 11 |
| 3.    | Abast.....                                | 12 |
| 3.1   | Objectius .....                           | 12 |
| 3.2   | Subobjectius i indicadors.....            | 12 |
| 3.3   | Requeriments.....                         | 13 |
| 3.4   | Obstacles i riscos .....                  | 14 |
| 4.    | Metodologia .....                         | 15 |
| 4.1   | Metodologia de treball.....               | 15 |
| 4.2   | Seguiment .....                           | 16 |
| 5.    | Planificació temporal.....                | 17 |
| 5.1   | Descripció de les tasques .....           | 17 |
|       | GP - Gestió del Projecte .....            | 17 |
|       | DA – Desenvolupament de l’algoritme ..... | 19 |
|       | AA – Aplicació de l’algoritme .....       | 22 |
| 5.2   | Recursos .....                            | 23 |
| 5.3   | Diagrama de Gantt .....                   | 25 |
| 5.4   | Gestió del risc.....                      | 27 |
| 5.5   | Canvis a la planificació inicial .....    | 28 |
| 5.5.1 | Problema .....                            | 28 |
| 5.5.2 | Solució decidida.....                     | 29 |
| 5.5.3 | Diagrama de Gantt .....                   | 30 |
| 6.    | Gestió Econòmica.....                     | 32 |
| 6.1   | Pressupost .....                          | 32 |
| 6.1.1 | Recursos humans .....                     | 32 |
| 6.1.2 | Hardware.....                             | 33 |
| 6.1.3 | Software .....                            | 34 |

|        |   |    |
|--------|---|----|
| 6.1.4  | Costs generals .....                              | 35 |
| 6.1.5  | Contingència.....                                 | 35 |
| 6.1.6  | Imprevistos.....                                  | 35 |
| 6.1.7  | Cost total .....                                  | 36 |
| 6.2    | Control de gestió .....                           | 37 |
| 7.     | Sostenibilitat i compromís social.....            | 37 |
| 7.1    | Autoavaluació.....                                | 38 |
| 7.2    | Estudi de l'impacte econòmic .....                | 38 |
| 7.3    | Estudi de l'impacte econòmic .....                | 39 |
| 7.4    | Estudi de l'impacte ambiental .....               | 40 |
| 7.5    | Matriu de sostenibilitat .....                    | 41 |
| 8.     | Marc legal.....                                   | 41 |
| 9.     | Disseny de la solució .....                       | 42 |
| 9.1    | Anàlisi de les eines .....                        | 42 |
| 9.1.1  | Llibreria NLTK .....                              | 43 |
| 9.1.2  | Snowball .....                                    | 44 |
| 9.1.3  | Llibreria Tensorflow.....                         | 44 |
| 9.1.4  | Llibreria Scikit-Learn .....                      | 44 |
| 9.1.5  | Llibreria Numpy i Pandas.....                     | 45 |
| 9.1.6  | Llibreria Matplotlib.....                         | 45 |
| 10.    | Implementació .....                               | 45 |
| 10.1   | Dataset: .....                                    | 46 |
| 10.1.1 | Opcions plantejades.....                          | 46 |
| 10.1.2 | Opció escollida .....                             | 47 |
| 10.1.3 | Etiquetatge del data set .....                    | 47 |
| 10.1.4 | Repercussions dels diferents intents .....        | 48 |
| 10.2   | Pre processat de les dades:.....                  | 48 |
| 10.2.1 | Tokenització.....                                 | 48 |
| 10.2.2 | Eliminació de stopwords .....                     | 49 |
| 10.2.3 | Puntuació .....                                   | 49 |
| 10.2.4 | Uppercase & lowercase.....                        | 50 |
| 10.2.5 | Stemming & lematizaci3n .....                     | 50 |
| 10.2.6 | Count vectorize .....                             | 50 |
| 10.3   | Generaci3n del model .....                        | 51 |
| 10.3.1 | Creaci3n del nostre propi classificador .....     | 51 |
| 10.3.2 | Classificadors de la llibreria scikit-learn ..... | 52 |

|   |    |
|---|----|
| 11. Resultats: avaluació i optimitzacions .....       | 54 |
| 11.1 Mètriques d'avaluació:.....                      | 54 |
| 11.2 Procés d'avaluació:.....                         | 54 |
| 11.3 Anàlisi classificador escollit: .....            | 56 |
| 12. Conclusions.....                                  | 57 |
| 12.1 Problemes que han sorgit .....                   | 58 |
| 12.2 Futures millores i continuació del projecte..... | 58 |
| Bibliografia .....                                    | 59 |
| Referències GEP: .....                                | 59 |
| Sobre el Dataset: .....                               | 60 |
| Sobre el Model: .....                                 | 61 |
| Sobre les Llibreries: .....                           | 62 |
| Annexes .....   | 64 |
| Annex A: Stopwords .....                              | 64 |

## Índex de Figures

|  |    |
|--|----|
| Figura 1: Diferències entre deep learning i Machine learning .....                               | 9  |
| Figura 2: Taula resum de les tasques, les seves dependències i hores de dedicació estimades..... | 25 |
| Figura 3: Diagrama de Gantt del projecte .....   | 26 |
| Figura 4: Taula de llistat de tasques del projecte actualitzada .....                            | 30 |
| Figura 5: Diagrama de Gantt actualitzat .....  | 31 |
| Figura 6: Taula de sous de creació pròpia basada en la guia salarial de Hays .....               | 32 |
| Figura 7: Taula amb els costos estimats de les tasques de la planificació temporal .....         | 33 |
| Figura 8: Taula de costos estimats de hardware.....  | 34 |
| Figura 9: Taula de costos software estimats .....  | 34 |
| Figura 10: Taula dels costos generals estimats .....   | 35 |
| Figura 11: Taula dels costos estimats amb un 15% de contingència .....                           | 35 |
| Figura 12: Taula del cost afegit per imprevistos .....   | 36 |
| Figura 13: Taula de pressupost final estimat del projecte .....                                  | 36 |
| Figura 14: Matriu de sostenibilitat .....  | 41 |
| Figura 15: Taula d'eines usades en el projecte .....   | 43 |
| Figura 16: Exemple d'una oració inicial que tenim en el dataset .....                            | 48 |
| Figura 17: Exemple oració del data set sense les stopwords.....                                  | 49 |
| Figura 18: Exemple oració del data set sense els signes de puntuació.....                        | 49 |
| Figura 19: Exemple oració del data set tot en minúscules .....                                   | 50 |
| Figura 20: Exemple oració després stemming i lemmatization .....                                 | 50 |
| Figura 21: Exemple taula final de la oració. ....  | 51 |
| Figura 22: Resum del model seqüencial implementat.....   | 52 |
| Figura 23: Gràfica de la accuracy en l'entrenament del model .....                               | 55 |
| Figura 24: Gràfica comparativa de la accuracy dels diferents models.....                         | 55 |

|  |    |
|--|----|
| Figura 25: Taula amb els valors de les mètriques analitzades ..... | 56 |
| Figura 26: Corba ROC del model Kneighbors escollit.....            | 56 |
| Figura 27: Matriu de confusió del model Kneighbors escollit.....   | 56 |

## Índex d'abreviatures

ACC: Accuracy

CPU: Unitat Central de Processament

FN: Fals Negatiu

FP: Fals Positiu

GPU: Unitat de Processat Gràfic

iSOCO: Organització Internacional de Comissions de Valors

MLP: MultiLayer Perceptron

NLP : Natural Language Processing

NLTK: Natural Language Tool Kit

PLN: Processat del Llenguatge Natural

ROC : Característica Operativa del Receptor

SKLEARN: SciKit-Learn

TN: True Negatives

TP: True Positives

VM: Virtual Machine

## 1. Introducció i contextualització

La Seguretat Informàtica és un àmbit que compren molts àmbits i que busca la seguretat en els sistemes informàtics per part dels usuaris. Aquesta compren nombroses mesures de seguretat que donen solució a la gran varietat d'atacs que es produeixen a les xarxes, entre aquests, és troba la variant del ciberbullying.

El ciberbullying és una de les situacions d'assetjament que més és troben en la actualitat. Segons l'estudi realitzat per la Fundació ANAR i la fundació Mutua Madrileña<sup>1</sup>, el ciberbullying representa un de quatre casos d'assetjament escolar a Espanya, xifra que empitjora a partir dels 13 anys on el 36.5% dels casos d'assetjament que es produeixen (més d'un de cada tres) són ciberbullying<sup>2</sup>.

Aquests casos son molt difícils d'evitar, en part per la habilitat de crear múltiples identitats en les xarxes socials, de manera que les eines de bloqueig de perfils que moltes plataformes ofereixen, no son efectives en aquests atacs, que es solen prolongar al temps (el 40% els casos estudiats portaven entre un mes i un any succeint<sup>3</sup>) i que si bloqueges un perfil n'apareixen múltiples que protegeixen en molts casos la identitat del atacant deixant-lo en l'anonimat i produint una situació d'indefensió a la víctima.

### 1.1 Contextualització

Aquest projecte entra dins el context de la menció de Tecnologies de la Informació, la qual és una de les cinc mencions a les que s'opta en el Grau d'Enginyeria Informàtica impartit per la Facultat D'Informàtica de Barcelona, una de les facultats que conformen la Universitat Politècnica de Catalunya.

En concret, parlem de l'àmbit de la seguretat informàtica aplicat a les aplicacions i busquem aplicar de manera principal la competència tècnica de l'especialitat de *concebre sistemes, aplicacions i serveis basats en tecnologies en xarxa, tenint en compte Internet, web, comerç electrònic, multimèdia, serveis interactius i computació ubiqua*<sup>4</sup>, d'entre altres.

El projecte està realitzat de manera autònoma sense formar part de ninguna empresa o equip d'investigació, i sota les guies del director del treball, que està especialitzat en el àmbit de la seguretat informàtica i ha suggerit l'idea sobre la que es basa el concepte del projecte.

---

<sup>1</sup> De Castañeda, A. (2018). *El ciberbullying en España*. Zonamovilidad.es. Retrieved February 2021, from <https://www.zonamovilidad.es/ciberbullying-en-espana-iii-informe-fundacion-anar-fundacion-mutua-madrilena>.

<sup>2 3</sup> *Datos sobre Bullying y Ciberbullying o acoso digital en España*. Blog Educación y Bienestar digital. (2017). Retrieved 2 March 2021, from <https://gaptain.com/blog/bullying-ciberbullying-acoso-espana/>.

<sup>4</sup> *Tecnologies de la informació | Facultat d'Informàtica de Barcelona*. Fib.upc.edu. (2021). Retrieved February 2021, from <https://www.fib.upc.edu/ca/estudis/graus/grau-en-enginyeria-informatica/pla-destudis/especialitats/tecnologies-de-la-informacio>.

## 1.2 Definició de conceptes

A continuació es defineixen els conceptes claus per a poder entendre el projecte i el plantejament d'aquest.

### 1.2.2 Cyberbullying

El cyberbullying forma el part de l'àmbit de la seguretat informàtica i és un concepte que defineix les accions que es duen a terme mitjançant mitjans informàtics per a l'assetjament, abús de forma repetida i sostinguda per part d'un o un grup d'individus a una persona.

La característica distintiva per a produir-se es l'ús de xarxes informàtiques i els recursos tecnològics actuals i és manifesten en espais com xats, fòrums, correus electrònics, xarxes socials o de manera generalitzada, qualsevol medi que permeti la interacció entre diferents individus.

Les conseqüències del cyberbullying poden resultar devastadores per als individus que son assetjats ja que afecten directament al seu benestar psicològic i emocional, fent-los més susceptibles a trastorns psicològics com la depressió i ansietat i arribant a comportar accions greus com l'autolesió.

### 1.2.2 Deep learning

El deep learning és un aspecte de la intel·ligència artificial que té l'objectiu de simular la forma d'aprenentatge que tenim els humans per a obtenir certs tipus de coneixements. Aquest tipus d'aprenentatge automàtic permet que mitjançant múltiples capes de càlcul i característiques d'alt nivell i extretes per la màquina es puguin resoldre problemes de gran complexitat.

Els algorismes de deep learning difereixen dels algorismes de Machine learning d'entre altres motius, per la extracció de les característiques. En el primer tipus, la extracció és produïx automàticament per la màquina, en caixes negres, és a dir, el programador no sap les característiques que la màquina obté de les dades rebudes, en canvi en Machine learning, les característiques sobre les que treballarà la màquina per a resoldre el problema, venen donades per el programador.

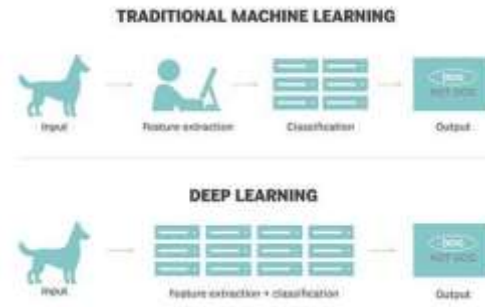


Figura 1: Diferències entre deep learning i Machine learning<sup>5</sup>

### 1.2.3 Processat del llenguatge Natural

El processat del llenguatge natural (PLN)<sup>6</sup> és un dels camps al que s’hi pot aplicar els algorismes de deep learning, i s’encarrega d’investigar maneres de comunicar les màquines amb les persones mitjançant el ús i tractament de les llengües naturals.

El llenguatge humà no és rígid i les paraules no sempre tenen les mateixes connotacions i significats sinó que aquests poden variar segons el context en el que s’escriguin el que pot comportar una dificultat de la comprensió al xocar amb la inflexibilitat de les màquines. El camp del processat del llenguatge natural busca canviar i desenvolupar aquesta relació de manera que les màquines puguin treballar amb unes dades tan versàtils com són les llengües naturals.

### 1.3 Identificació del problema

Actualment existeixen milions de diferents maneres per a interactuar entre individus i per tant per a que es produeixin situacions de ciberbullying que poden ser molt difícils d’evitar no només per la possibilitat d’anonimat que et dona la interacció per internet sinó també per la manca d’interacció en persona, el que fa difícil de detectar aquestes situacions i assegurar una intervenció d’algú per a ajudar a parar el ciberbullying.

Per a intentar impedir aquesta problemàtica, aquest projecte busca desenvolupar un plugin que permeti que mitjançant un algorisme de deep learning, alimentat per textos i paraules que reflecteixen el llenguatge i argot actualment més utilitzat, permeti detectar la intenció dels textos que es reben i determinar si són textos amb intenció maliciosa, és a dir, que amenacin, o faltin al respecte, per a poder realitzar accions sobre aquests textos amb l’objectiu d’ evitar el ciberbullying a les xarxes.

<sup>5</sup> ¿Qué es Aprendizaje profundo (deep learning)? - Definición en Whatls.com. SearchDataCenter&nbsp;en&nbsp;Español. (2017). Retrieved February 2021, from <https://searchdatacenter.techtarget.com/es/definicion/Aprendizaje-profundo-deep-learning>

<sup>6</sup> Moreno, A. *Procesamiento del lenguaje natural ¿qué es?* - IIC. Instituto de Ingeniería del Conocimiento. Retrieved February 2021, from <https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/>.

## 1.4 Actors Implicats

El projecte va dirigit a tots els usuaris que interaccionin amb altres utilitzant les xarxes informàtiques, però en particular a la gent jove i menors, que són el col·lectiu més susceptible als atacs de ciberbullying als quals el nostre projecte busca donar-li solució. En una situació ideal aquest algoritme aplicaria a tot el contingut web eliminant el així el contingut agressiu que s'hi pot trobar i fent de beneficiaris a tots els usuaris els quals es beneficiarien d'un entorn molt més segur i positiu.

Per tant, l'algoritme que proposa aquest treball pot ser usat per usuaris de l'Internet que vulguin filtrar els missatges dels quals en son part per evitar ser els receptors de conductes agressives i amenaçants, així com propietaris de blogs i xarxes socials que vulguin que la seva plataforma sigui un espai més segur des del punt de vista de l'adequació de les interaccions socials que es duen a terme.

Així doncs, els actors que trobem per el desenvolupament d'aquest projecte son :

- *El personal del projecte:* En aquest grup entra l'autora del projecte que serà la encarregada del desenvolupament d'aquest i els diferents tutors de la Facultat d'Informàtica assignats a aquest treball que l'orientaran i guiaran, com el director del projecte i el tutor del GEP assignat a aquest treball.
- *Els usuaris d'Internet especialment menors d'edat:* Ja que l'objectiu d'aquests projecte es desenvolupar una eina que permeti detectar texts amenaçants de manera que es pugui protegir contra el ciberbullying als usuaris d'internet. Especialment persones menors d'edat, doncs són les principals víctimes d'aquests atacs. Aquests per tant seran els majors beneficiaris d'aquest projecte i els que l'usaran més.
- *Escoles i xarxes socials:* Amb l'objectiu d'aplicar aquesta eina per a defensar els menors, els principals actors que en farien ús, deixant de banda l'ús individual dels propis menors per a protegir-se, serien les escoles per a fer les interaccions que succeeixen dins l'entorn escolar mitjançant medis informàtics segurs, i també les xarxes socials, doncs és on succeeixen la gran majoria d'aquests atacs.

## 2. Justificació

### 2.1 Situació actual

Un dels grans problemes que ens trobem al moment d'intentar enfrontar els atacs de ciberbullying són el gran nombre de llengües que existeixen en aquest món. Segons la revista

*Ethnologue*<sup>7</sup>, es calcula que existeixen 7097<sup>8</sup> llengües arran del món, el que son set mil idiomes amb els que realitzar possibles amenaces a diferents usuaris.

Aquest projecte busca donar solució a el ciberbullying de manera concreta, deixant de banda altres eines existents com són el bloqueig de perfils i que no arriben a eradicar aquestes situacions. Per a fer-ho s'enfoca amb un algorisme dins del processat del llenguatge natural i per conseqüència hem de decidir quin llenguatge treballarem, aquest s'ha decidit que serà el català, doncs és la llengua materna de la autora.

## 2.2 Solucions ja existents

Existeixen diferents eines que utilitzant el processat del llenguatge natural busquen analitzar els missatges rebuts i dur a terme una aplicació similar a la que proposa aquest projecte. Entre d'altres dignes de menció i que han servit d'inspiració hi ha:

- *Bullstop*<sup>9</sup>  
És una aplicació en anglès que revisa els missatges rebuts en les diferents xarxes socials i marca aquells que tenen contingut ofensiu de bullying, abús, insults, etc, i et permet bloquejar els comptes que envien els missatges o eliminar-los de manera automàtica sense que els tinguis que llegir i interactuar tu amb ells.
- *Jimcrick*<sup>10</sup>  
Eina en castellà que detecta els missatges ofensius i mitjançant un algorisme intern determina si l'usuari està patint ciberbullying. En cas de que estigui succeint s'envia un missatge a un compte determinat, de manera que un adult s'assabenti de la situació i pugui prendre les accions que cregui corresponent sense envair la intimitat del usuari que usa la aplicació llegint els seus missatges i interaccions digitals.

## 2.3 Solució decidida

Un dels grans inconvenients del processat del llenguatge natural és que el algorisme s'ha d'entrenar amb un data set del llenguatge que entindrà i podrà processar. De les eines descobertes ja existents, no hem trobat cap que apliqui el processat del llenguatge natural amb el català de manera que aquest es un sector que encara no ha rebut una solució d'aquest tipus aquí a Catalunya.

---

<sup>7</sup> *Ethnologue: Languages of the World*. Ethnologue. Retrieved February 2021, from <https://www.ethnologue.com/>.

<sup>8</sup> *Los idiomas, en cifras: ¿cuántas lenguas hay en el mundo?*. europapress.es. Retrieved February 2021, from <https://www.europapress.es/sociedad/noticia-idiomas-cifras-cuantas-lenguas-hay-mundo-20190221115202.html>.

<sup>9</sup> *BullStop - Welcome to a safer Internet*. Bullstop.io. Retrieved February 2021, from <https://www.bullstop.io/>.

<sup>10</sup> *Jimcrickapp*. <https://jimcrickapp.com/>. (2021). Retrieved February 2021, from <https://jimcrickapp.com/>.

Per altra part les eines que hem mencionat i trobat que existien i treballaven un entorn similar a aquest projecte, van enfocades a que l'usuari d'elles sigui l'usuari d'internet que es protegeix a si mateix, però no ofereixen una solució als creadors de contingut per a fer més segures les seves plataformes.

Per tant, el projecte presenta una idea ja explorada però contextualitzada de manera diferent, ja que facilitaria també als autors de blogs o propietaris de xarxes socials a que apliquessin aquesta eina a les seves plataformes de manera que poguessin implementar aquest nivell addicional de seguretat en la xarxa, a més de fer-ho en un llenguatge que encara no s'ha explorat.

### 3. Abast

Al ser un projecte amb temps de desenvolupament limitat, busquem definir l'abast d'aquest determinat els objectius i requeriments per a considerar el projecte satisfactori.

#### 3.1 Objectius

L'objectiu principal d'aquest treball és el disseny, desenvolupament i implementació d'un algoritme de deep learning que mitjançant les eines del processat del llenguatge natural et permeti analitzar un text i extrapolar-ne la intenció d'aquest. Com a objectiu secundari posaríem aplicar aquest algoritme mitjançant alguna eina, com un plugin a alguna plataforma (com un blog) per a arribar a mostrar un ús real i realitzar accions com a conseqüència del resultat obtingut al processar texts amb l'algoritme desenvolupat. Aquest segon objectiu vindrà però limitat per el requisit essencial de finalitzar el treball de fi de grau una setmana abans de la seva exposició, com marquen els límits establerts per la Facultat d'Informàtica de Barcelona.

#### 3.2 Subobjectius i indicadors

Els subobjectius i indicadors en els que es dividirà i que són definits en el projecte per ordre de precedència i consisteixen en els següents:

##### 1. Generació de l'algoritme

###### 1. Creació d'un diccionari com a data set

- a) Creem un data set amb texts comuns usats en atacs de ciberbullying i amb texts que reflecteixin tant llenguatge correcte com llenguatge més col·loquial i complet.
- b) Etiquetatge dels text del data set de manera positiva o negativa segons el missatge
- c) Avaluació del data set perquè sigui equitatiu per evitar esbiaixats.

###### 2. Programar l'algoritme de deep learning

- a) Creació de l'algoritme

- b) Avaluació estadística del funcionament d'aquest
- 3. Entrenament de l'algoritme**
  - a) Entrenar el algoritme amb el data set corresponent
- 4. Avaluat l'algoritme**
  - a) Determinar un llindar d'acceptació dels resultats
  - b) Avaluat els resultats obtinguts de forma estadística
  - c) Repetir al pas tres si no s'arriba al llindar determinat d'encerts

## 2. Aplicació

Com hem mencionat de manera prèvia en l'apartat 3.1, el camp i l'eina que usarem per aplicar l'algoritme esta subjecte a variacions segons el temps amb el que comptem, així com tota l'aplicació del segon objectiu. A continuació definim els subobjectius que realitzaríem si es pogués dur a terme tot el treball de forma ideal.

- 1. Generar un plugin de wordpress<sup>11</sup>**
  - a. El plugin registra texts de les plataformes
  - b. El plugin envia els text al algoritme generat i obté un resultat
  - c. El plugin realitza una acció segons el resultat obtingut
- 2. Aplicació a una plataforma**
  - a. Creació d'una plataforma per a fer la prova (coneixent així les especificacions de dita plataforma)
  - b. Aplicació del plugin a aquesta plataforma

## 3. Avaluació

Finalment com a tercer objectiu del projecte trobaríem l'avaluació de la nostra solució implementada en contraposició amb altres solucions ja existents per a determinar la viabilitat i encert de la solució a la que hem arribat amb les condicions a les que ens trobàvem i si difereix de la solució idealment plantejada en el projecte, per què.

### 3.3 Requeriments

Els requeriments que trobem en el projecte que desenvoluparem, són els especificats a continuació:

1. És necessari fer una investigació prèvia de la manera en que es desenvolupen els atacs de cyberbullying, i el tipus de llenguatge emprat per a poder desenvolupar un diccionari que sigui complet i actualitzat amb el argot col·loquial, de manera que s'adapti al llenguatge emprat avui en dia.
2. És necessari definir un mètode de puntuació del valor dels texts de manera que l'aprenentatge del algorisme sigui correcte i equitatiu sense esbiaixats.

---

<sup>11</sup> *WordPress.com: crea un sitio web o blog gratuito.* WordPress.com. Retrieved February 2021, from <https://wordpress.com/es/>.

3. És necessari determinar la mida del data set a generar així com assegurar que el contingut d'aquest sigui suficient i adient.
4. És necessari determinar els paràmetres i característiques que avaluaran el funcionament de l'algorisme de manera objectiva i justificada.
5. Ens convé conèixer de les especificacions de la plataforma on s'aplicarà l'eina desenvolupada per a poder tenir un correcte funcionament en aquesta.
6. És indispensable no fer cap mena d'emmagatzematge dels texts registrats per a no envair la privacitat de l'usuari.

## 3.4 Obstacles i riscos

### 3.4.1 Obstacles

Hem determinat quatre obstacles principals en el correcte desenvolupament del projecte, llistats a continuació:

#### 1. *Complexitat del data set*

L'algorisme que es vol desenvolupar requereix d'un data set amb un alt grau de complexitat amb paraules i texts avaluats objectius equitatius i actuals. La dificultat en la recerca de l'argot usat actualment, la elaboració d'una llista de paraules suficientment llarga i del gran nombre de texts a ser avaluats per a que el data set sigui correcte i suficient per a un bon resultat de l'algorisme pot requerir d'uns coneixements i quantitat de temps de la que no es disposen i que pot complicar greument el desenvolupament del projecte.

#### 2. *Desconeixement de la arquitectura*

L'eina que busquem desenvolupar, de manera ideal es podria aplicar a qualsevol plataforma o xarxa social existent. Aquesta però, necessita fer un hàbil i correcte enregistrament dels texts que es continguin. Les diferents maneres en que les plataformes exposin el text pot fer que aquest no es pugui detectar o que no es puguin realitzar les accions desitjades per manca de coneixement del funcionament intern de cada plataforma, així com les diferents especificacions d'aquestes. Com a solució es mira d'implementar la nostra pròpia plataforma de manera senzilla tot i que això també comportaria un gran cost d'un temps ja limitat.

#### 3. *Inexperiència amb les eines emprades*

L'eina que es desitja implementar consisteix en un plugin de wordpress, doncs es considerarà que es la manera més adient per a la solució que es vol implementar. L'autora d'aquest treball té però una experiència molt limitada i petita amb wordpress, el que li suposarà un esforç extra i pot comportar problemes imprevistos per manca de coneixement.

#### 4. *Equipament*

El projecte serà desenvolupat en el entorn local de l'autora que desenvoluparà tot l'algoritme i realitzarà totes les proves amb el seu ordinador personal i amb les limitacions de CPU i capacitat que això comporta.

#### 3.4.2 *Riscs*

S'han determinat tres riscs principals en el enfoc del projecte que poden dificultar el seu correcte desenvolupament:

##### 1. *Esbiaixat del data set*

Com hem especificat en l'apartat 3.1.1 les grans dificultats de generació del data set per a entrenar el algoritme poden fer que aquest no sigui suficientment bo i afectar els resultats considerablement, e inclús generar un esbiaixat que eviti que les etiquetes obtingudes per el algoritme no siguin les correctes, el que posaria en perill tot el projecte.

##### 2. *Imprevistos pel desconeixement del sistema*

Tant el desconeixement de les particularitats de les arquitectures sobre on s'aplicaria el plugin tal com el wordpress amb el que es generaria poden generar greus imprevistos sobre el treball tant com que per manca d'habilitats no s'aconsegueixi que l'eina faci tot això que es desitjaria com perquè no es pugui aplicar a segons quines plataformes dependent de les particularitats de les seves arquitectures.

##### 3. *Gestió del temps*

Degut al desconeixement de molts factors del treball, especificats prèviament i mencionats altra vegada en aquest mateix apartat en el punt dos, poden sorgir imprevistos que no permetin que es segueixi el pla de treball tal i com estava definit. Sumat això a altres possibles imprevistos en les diverses activitats lectives que cursa l'autora del TFG simultàniament, pot comportar el risc de no aconseguir completar tots els objectius tal qual estan pensats de forma ideal o hauré d'adaptar el projecte de manera que es pugui finalitzar en el termini indicat.

## 4. Metodologia

### 4.1 Metodologia de treball

Aquest Treball de Final de Grau, com ha estat explicat en l'apartat 3.1.2 té un gran risc a la gestió de temps doncs la manca de experiència amb les eines i en el temari que es treballarà impliquen una falta de coneixement del temps i duració de les activitats. En conseqüència es necessita una metodologia que permeti la variabilitat i adaptació de les diferents tasques del projecte de

manera que s'obtingui un treball satisfactori per la data final estipulada per la Facultat d'Informàtica de Barcelona.

En concret, la metodologia que es farà servir serà una metodologia àgil, on en primer lloc es definirà un pla temporal de treball i després de manera periòdica s'aniran realitzant reunions amb el director de projecte per a determinar si els objectius marcats s'han complert i si els terminis son satisfets, o si, pel contrari, ens hem trobat en alguna situació que requereix de la variació del pla temporal del treball o fins i tot dels objectius definits en el projecte.

Aquestes reunions serviran també per afegir un altre punt de vista objectiu a la validació de les tasques completades, les quals seran provades per jocs de proves i es comprovarà que satisfacin un llinar d'encert abans de declarar-les com a exitoses.

A més, les tasques determinades en el pla temporal estan fetes de manera seqüencial i per ordre de prioritat i rellevància en el projecte. Amb això ens referim a que per assegurar que es completa el projecte en el termini requerit, primer es duran a terme parts del projecte essencials i una vegada completades, s'avançarà a altres tasques derivades i que poden, en cas extrem, ser eliminades amb una reducció del abast en cas de ser impossible de dur a terme a temps.

A mesura que es vagin finalitzant les tasques definides, s'anirà comprovant que el pla temporal dissenyat s'estigui complint. En cas contrari es prendran decisions per a rectificar el desviament del desenvolupament del treball. Aquestes decisions correctives i possibles obstacles venen explicats amb detall al punt 5 sobre la planificació temporal.

## 4.2 Seguiment

Com hem anat explicant al llarg del document, els objectius poden ser variants segons el ritme del desenvolupament del projecte, és per això molt important determinar unes eines per a realitzar el seguiment dels objectius i el temps emprat de manera que puguem realitzar els canvis adients a temps per a finalitzar el projecte de forma exitosa.

Principalment farem ús de dos eines. La primera és l'eina *Ganttter*<sup>12</sup>. Aquesta eina ens serveix per a realitzar la planificació de tasques del projecte. Amb la creació d'un diagrama Gantt, realitzarem la planificació de tasques temporalment de manera inicial fent una planificació per a finalitzar el projecte dins del termini marcat, i al llarg del desenvolupament del projecte segons l'avenç real dut a terme, s'anirà adaptant perquè realitzi un control fiable del temps.

La segona eina que usarem serà *Trello*<sup>13</sup>. Aquesta es una eina que permet definir les diferents tasques en taulers virtuals de manera que puguem definir les diferents tasques a dur a terme en els diferents períodes de temps i portar així un registre i control de les tasques realitzades i realitzar de manera intuïtiva i visual.

---

<sup>12</sup> *Ganttter* | #1 Cloud-Based Project Management Software. Ganttter. Retrieved February 2021, from <https://www.ganttter.com/>.

<sup>13</sup> *Trello*. Trello.com. Retrieved February 2021, from <https://trello.com/>.

## 5. Planificació temporal

Un dels requeriments i objectius principals del treball de fi de grau és el d'acabar dins del termini especificat per la Facultat d'Informàtica de Barcelona, sota la qual és realitza el projecte. Per a aconseguir-ho especificuem una planificació temporal de les tasques a dur a terme per a finalitzar a temps.

Aquest treball de fi de grau comença el Dimarts 23 de febrer del 2021 i es preveu finalitzar-lo el 20 de Juny del 2021. Per tant és calcula que el desenvolupament del treball es durà a terme en 18 setmanes. La 18 ocupant tasques de preparació de la presentació però amb la documentació ja acabada i entregada de cara a la defensa del projecte.

És calculà que és tindrà una dedicació diària de 4 hores al projecte, el que comporten unes 28 hores setmanals i amb una estimació de 504 hores totals. Amb l'objectiu de mantenir una eficiència en el desenvolupament del treball, s'intentarien compensar les hores que no s'hagin pogut dedicar de dilluns a divendres per qualsevol imprevist, allargant les hores dedicades en el cap de setmana on es hi ha menys compromisos i més oportunitat de dedicació.

### 5.1 Descripció de les tasques

En aquesta secció detallem de forma individual les tasques en les que es dividirà el treball per a complir els objectius del projecte. Aquestes tasques formen part de diferents blocs segons la naturalesa d'aquestes de manera que és faciliti la comprensió de les diferents fases del treball que es dura a terme. Per a cada tasca que s'especifica a continuació és donarà una breu explicació de que consisteix la tasca, les hores de dedicació d'aquestes i si té alguna dependència amb altres tasques que s'han de dura a terme.

#### GP - Gestió del Projecte

La gestió del projecte és un àmbit molt necessari en el treball doncs és on es duen a terme les tasques de planificació del treball i del temps en el que es durà a terme, especificacions i definicions del treball a realitzar, i la documentació d'aquest. La gestió del projecte es realitza de manera continua al llarg del desenvolupament de tot el treball i de manera progressiva i estimem que de manera global, al llarg de les 18 setmanes que dura el projecte, donarem a les tasques de gestió del projecte una dedicació de 180 hores.

##### GP.1 – Abast

De les primeres tasques a dur a terme en la realització d'un projecte és definir l'abast d'aquest. Per tant aquesta tasca és de les inicials del projecte i ens ajuda a definir quin es l'objectiu del

treball de manera concreta, amb quines eines comptem i necessitem per a dur-lo a terme i quins son els objectius i subobjectius a complir que determinaran que el projecte ha sigut un èxit.

Per a dur-lo a terme cal prèviament haver fet una investigació del temari i eines que es tractaran, tindre unes nocions bàsiques dels conceptes amb els que es treballaran i complir els requeriments de la Facultat d'Informàtica de registrar el treball, tenir un director disposat a portar-te'l i que aquest sigui aprovat per la facultat.

En total s'estima una dedicació de 28 hores que inclouen tot el treball previ i la realització de la primera entrega en el GEP.

#### *GP.2 – Planificació*

La planificació temporal del projecte és una de les tasques requerides per a assegurar un bon funcionament del desenvolupament del treball al llarg del temps i assegurar la realització satisfactòria d'aquest en el termini indicat. Aquesta tasca consisteix en la especificació de les diferents tasques a dur a terme en el projecte i la dedicació que requereixen aquestes per a poder fer una estimació del temps necessari per a completar tots els objectius i permetre tindre un control del avanç del projecte a mesura que avança el temps que permeti l'ajust en cas de dificultats.

La planificació temporal del projecte requereix la definició prèvia de l'abast del projecte i l'anàlisi dels possibles riscos i obstacles amb els que ens podem topa en el desenvolupament d'aquest per a poder generar els plans i accions de rectificació requerits per a poder completar el projecte satisfactòriament.

S'estima una dedicació de 24 hores que inclouen la segona entrega del GEP.

#### *GP.3 – Gestió econòmica i de sostenibilitat*

En aquesta tasca es documenten el pressupost definit per a dur a terme el projecte, el cost d'aquest i la sostenibilitat del treball.

Aquesta definició del cost del projecte depèn de la definició del abast del projecte, recursos necessaris i planificació temporal doncs si el treball compta amb treballadors, el temps d'aquests te un cost. A més s'ha de realitzar un anàlisi del impacte mediambiental, social i econòmic del projecte.

S'estima una dedicació necessària a aquesta tasca de 20 hores, que inclouen la tercera entrega del GEP.

#### *GP.4 – Documentació fase inicial*

Aquesta tasca inclou l'agrupació de les tres tasques prèviament entregades del GEP en un document ben format i que inclogui les modificacions pertinents una vegada rebuda la retroacció de les entregues prèvies realitzades.

Com el document compren les entregues anteriors, aquesta tasca depèn de les tres primeres dutes a terme: GP.1, GP.2, i GP.3 i de les correccions realitzades per el tutor de GEP assignat a aquest projecte.

S'estima una dedicació de 5 hores per a dur a terme la quarta i última entrega de GEP.

#### *GP.5 – Reunions de seguiment*

Al llarg del desenvolupament del projecte s'han de dur a terme diferents reunions amb el director del treball per analitzar el progrés que s'està tenint, i si s'està seguint la planificació creada inicialment o si per altra banda ens hem topat amb algun obstacle que ens obliga a realitzar canvis per a poder finalitzar el treball satisfactòriament.

Amb la situació actual de la Covid-19, la majoria de les comunicacions es duran a terme de manera telemàtica per evitar al màxim contactes innecessaris de manera que es preveuen interaccions setmanals per e-mail, i reunions una vegada al bisetmanals per analitzar el seguiment adient del projecte.

Estimem en total unes 15 hores realitzant aquestes activitats de seguiment del projecte.

#### *GP.6 – Documentació*

El treball de fi de grau requereix d'una memòria a entregar al final del projecte i que ha de reflectir tot el desenvolupament d'aquest al llarg de les diferents fases. Per a fer-ho s'anirà realitzant la documentació de les diferents tasques desenvolupades al moment que es desenvolupen, de manera que aquesta documentació s'anirà generant progressivament amb el desenvolupament del treball. S'estima que aquesta tasca ocuparà unes 60 hores de feina

#### *GP.7 – Presentació*

Seguint els terminis marcats per la Facultat d'Informàtica, la documentació del projecte s'ha d'entregar una setmana abans de la presentació per la defensa del projecte. Per tant calculem que aquesta ultima setmana es prepararà el material necessari i guió i assaigs necessaris per a una presentació adient. Calculem que la feina es produirà al llarg de la ultima setmana doncs la presentació depèn del projecte realitzat i per tant necessita de tota la documentació acabada. S'estima que les hores dedicades aquesta tasca seran 28 hores

### *DA – Desenvolupament de l'algoritme*

#### *DA.1 – Creació d'un diccionari com a data set*

##### **1. Investigació**

És necessita fer una investigació prèvia sobre quin argot es vol incloure al data set per a fer un diccionari inclusiu i actual amb el llenguatge usat a més de les paraules de

diccionari. Decidir també els texts que es faran servir a més del nombre necessari d'elements al data set i intentar definir un contingut equitatiu i no esbiaixat.

Aquesta part serà molt complexa ja que és la que definirà el contingut amb el que s'alimentarà l'algoritme de deep learning. Una de les majors complexitats serà la de escollir els textos i paraules correctes per a formar part del data set ja que és una elecció entre una gran quantitat de contingut disponible. A això se li suma la dificultat de determinar una mida adient per a no sobre entrenar l'algoritme però suficient per a assegurar un correcte funcionament. Per a aconseguir un resultat positiu es farà l'anàlisi de estudis sobre el llenguatge del ciberbullying ja existents, així com l'examinació i possible recolzament en data sets prèviament creats per a algoritmes similars com el data set públic existent de la Universitat de Cambridge.

Deguda a la gran complexitat i la necessitat imprescindible d'aquesta tasca per el funcionament del projecte, se li estima una dedicació de 28 hores.

## 2. *Elaboració d'un data set diccionari*

Seguint les decisions preses a la investigació, en aquesta tasca creem el data set en el format adient per a poder ser usat en l'entrenament del algoritme a generar. Aquesta tasca és depenent de la anterior ja que posa en pràctica les decisions preses anteriorment i forma del data set a usar per l'algoritme. Com el major pes de la tasca esta a la part d'investigació, i aquesta consisteix en el dur a terme el data set, li estimem una dedicació de 38 hores.

## 3. *Elaboració mètode d'avaluació per l'etiquetatge*

Decidir les etiquetes que se li donaran al data set (positives o negatives) i quin mètode es farà servir per a determinar l'etiquetatge de cada mostra del data set.

Aquesta tasca també serà complexa ja que es busca determinar un sistema de puntuació que permeti al algoritme determinar si el text és positiu o negatiu. Aquest sentiment positiu o negatiu s'ha d'especificar encara més generant etiquetes més concretes com: amenaçant, depressiu, violent..., per a poder d'aquesta manera diferenciar els texts negatius per tristesa dels texts negatius per assetjament que és el que busca determinar l'eina. Aquest sistema de puntuació tindrà que ser a nivell de paraula però també a nivell global de frase, per ajudar a que l'algoritme entengui el context de les oracions, cosa que aporta gran dificultat a la tasca.

Deguda a la complexitat de la tasca i a la seva essencialitat, determinem una dedicació en hores de 28.

## 4. *Etiquetatge dels elements del data set*

Fer l'etiquetatge de tots els atributs del data sets segons les decisions preses en el punt tres. Aquesta tasca es realitza després de la tercera, doncs depèn del sistema de puntuació creat per a fer l'etiquetatge de tots els atributs que formen el data set. Determinem doncs que ens durà una dedicació de 20 hores.

## DA.2 – Programar l’algoritme de deep learning

### 1. Investigació

Per a crear l’algoritme de deep learning primer tindrem que prendre diferents decisions de com el volem fer. Entre d’altres s’haurà de determinar el tipus de capa que volem utilitzar, el nombre de capes que tindrà, el nombre de neurones d’entrada i sortida de cada capa, etc... , així com decisions més bàsiques sobre el llenguatge a usar i les llibreries a ser utilitzades.

Per a prendre totes aquestes decisions necessitarem de recerca en el àmbit de deep learning doncs al moment de realitzar el projecte l’autora no ha dut a terme més que un parell de programes molt simples en l’entorn de deep learning i és completament inexperienciada en l’àmbit del processament del llenguatge natural. S’espera necessitar del coneixement de les característiques del data set amb el que serà entrenat i saber doncs que tindrà d’entrada. Per a realitzar la tasca s’estima una duració de 28 hores.

### 2. Generació de l’algoritme

Aquesta tasca compren la creació de l’algoritme. Seguint la presa de decisions feta en la tasca del punt anterior, és buscarà generar l’algoritme de manera que sigui eficient i doni la resposta esperada de la manera esperada i acordada. Aquí també comprem l’entrenament del algoritme doncs és necessari per a obtenir el resultat de l’algoritme. S’estima que la duració d’aquesta activitat, tenint també en compte temps d’execució, serà de 42 hores.

## DA.3– Avaluació de l’algoritme (HORAS)

### 1. Definició de paràmetres

Una vegada obtingut l’algoritme haurem d’avaluar el seu funcionament. Per a fer-ho s’han de determinar uns llimars de rendiment acceptables, dintre dels quals es decidirà que l’algoritme té un correcte funcionament, i assumint que no s’aconseguiran un 100% d’encerts. Per a decidir aquests llimars es necessitarà un estudi d’aquest tipus d’algorismes, per a donar un numero coherent i acceptable. La duració d’aquesta tasca s’estima en 5 hores.

### 2. Avaluació

Aquesta tasca es realitza una vegada creats el algoritme i alimentat el data set, així com decidit els llimars de rendiment acceptables. Consisteix en fer córrer l’algoritme i examinar els resultats i determinar si aquests son correctes i per tant el algoritme te un bon funcionament d’acord a les característiques decidides, o si en canvi el funcionament no és el suficientment bo. En aquest cas s’haurà de retornar a la tasca de creació del data set o de la creació del algoritme i modificar los de manera que s’obtinguin uns resultats exitosos. S’estima que la duració d’aquesta tasca serà de 2 hores, tot i que és la possible desencadenant de endarreriments per la possible necessitat de repetició d’altres tasques.

## AA – Aplicació de l'algoritme

### AA.1 – Generar un plugin de wordpress

#### 1. *Investigació*

Al moment de realitzar el treball, la autora no té més que els coneixements mínims i una total inexperiència amb wordpress. Per tant és necessària de recerca sobre com és pot generar un plugin que incorpori l'algoritme produït. Costa estimar el temps d'aquesta tasca degut al desconeixement del àmbit, però s'estimen que unes 28 hores seran suficients per completar-la.

#### 2. *Generació del plugin*

Una vegada informats i amb els coneixements necessaris per treballar amb wordpress, duríem a terme la generació de dit plugin. Al igual que amb la tasca anterior, la aproximació d'hores costa de fer-se degut a la desconexença de la autora en l'àmbit i l'eina a usar, però s'estima que unes 28 hores bastaran per satisfactòriament completar la tasca.

### AA.2 – Aplicació en una plataforma

#### 1. *Investigació*

Una vegada generat el plugin aquest s'ha d'aplicar a alguna plataforma per a mirar el seu funcionament. Per a evitar problemes amb la manca de coneixement de l'arquitectura de xarxes socials inexistents, aquest projecte es planteja crear la seva pròpia plataforma on aplicar el plugin. En aquesta tasca és realitzarà labor d'investigació per decidir quina plataforma és més adient i els mètodes i coneixements necessaris per a crear-la i estima una duració de 28 hores.

#### 2. *Creació de la plataforma*

Aquesta tasca compren la creació de la plataforma en la que s'aplicarà el plugin creat. Depèn doncs, de la labor d'investigació prèviament feta, que definirà quina és la plataforma adient i les especificacions de com crear-la. La durabilitat d'aquesta tasca pot ser molt variant, però com és una eina per a mostrar l'èxit del algoritme, que és realment al finalitat del projecte, s'enfocarà de la manera més senzilla possible i s'estima que una dedicació de 28 hores serà suficient per a dur-la a terme.

#### 3. *Aplicació del plugin a la plataforma*

Una vegada realitzades les tasques anteriors, tindrem la plataforma creada i el plugin creat. En aquesta tasca, ens encarreguem de combinar-los per a obtenir els resultats esperats. Si les tasques d'investigació s'han realitzat correctament, s'espera que no hi

hagin grans dificultats per a la implementació del plugin en la plataforma creada, raó per la qual estimem una durabilitat de 14 hores per a realitzar la tasca.

### AA.3 – Avaluació del funcionament de l'eina

#### 1. Investigació dels criteris

Aquesta tasca d'investigació serveix per a determinar els criteris que declararan un èxit la implementació del nostre plugin amb l'algoritme en una plataforma, per a demostrar la seva viabilitat. S'estimen unes 5 hores per a dur-la a terme.

#### 2. Avaluació del funcionament

Una vegada determinats en la tasca anterior els criteris que marcaran la avaluació, realitzem aquesta tasca d'avaluació a la plataforma creada amb el plugin creat per a comprovar si el projecte s'ha finalitzat exitosament o si per altra banda es necessiten visitar algun punt anterior i fer algunes modificacions per a aconseguir els llistats d'èxit marcats. Aquesta tasca té una durabilitat estimada de 2 hores.

## 5.2 Recursos

### 5.2.1 Recursos humans

Podem determinar diferents rols en el desenvolupament del projecte, que si aquest fos desenvolupat a gran escala comportarien diferents persones i permetrien un major paral·lelisme en la realització del projecte .La documentació seria generada de manera conjunta, documentant cadascú les diferents tasques que duen a terme. En el cas d'aquest treball de fi de grau però, tots i cadascun dels rols definits a continuació seran duts a terme per l'única persona participant, que és la autora del projecte.

- *Cap del projecte (CP)*

Aquest s'encarrega de les tasques de planificació i gestió del projecte i dels recursos d'aquests, és el líder i en última instància el factor decisiu en la presa de decisions.

- *Investigador (I)*

S'encarrega de les tasques d'investigació i determina en conseqüència el disseny de les diferents parts del projecte segons les característiques determinades fruit de la investigació duta a terme.

- *Programador (P)*

S'encarrega d'implementar el sistema, per tant de totes les tasques de generació del data set, algoritme i de l'entrenament d'aquest.

- *Avaluador (A)*  
S'encarrega de les tasques de prova i avaluació del treball realitzat així com de la determinació dels diferents criteris i l'indars per a establir que el desenvolupament del treball ha estat correcte.

### 5.2.2 Recursos materials

El material del que es disposa per a realitzar el projecte és l'especificat a continuació. Aquest material s'utilitza a totes les tasques del projecte i és essencial per a dur-lo a terme.

- Un espai de treball amb una ràpida connexió a Internet
- L'ordinador personal amb un sistema operatiu Windows.
- De les eines gratuïtes de: wordpress, gantter, google collab i Trello

| <b>Id.</b>  | <b>Tasques</b>                                | <b>Temps</b> | <b>Dependència</b> | <b>Recursos</b> |
|-------------|---|--------------|--------------------|-----------------|
| <b>GP</b>   | <b>Gestió del projecte</b>                    | <b>180h</b>  | -                  |                 |
| GP.1        | Abast   | 28h          | -                  | CP              |
| GP.2        | Planificació                                  | 24h          | GP.1               | CP              |
| GP.3        | Gestió econòmica i de sostenibilitat          | 20h          | GP.1               | CP              |
| GP.4        | Documentació fase inicial                     | 5h           | GP.1,GP.2,GP.3     | CP              |
| GP.5        | Reunions de seguiment                         | 15h          | -                  | CP, I, P, A     |
| GP.6        | Documentació                                  | 60h          | -                  | CP, I, P, A     |
| GP.7        | Presentació                                   | 28h          | GP.4, GP.6         | CP              |
| <b>DA</b>   | <b>Desenvolupament de l'algoritme</b>         | <b>191h</b>  | <b>GP</b>          |                 |
| <b>DA.1</b> | <b>Creació d'un diccionari com a data set</b> | <b>114h</b>  | -                  | -               |
| DA.1.1      | Investigació                                  | 28h          | -                  | I               |
| DA.1.2      | Elaboració d'un data set diccionari           | 38h          | DA.1.1             | P               |
| DA.1.3      | Mètode d'avaluació per l'etiquetatge          | 28h          | DA.1.2             | I               |
| DA.1.4      | Etiquetatge dels elements del data set        | 20h          | DA.1.3             | P               |
| <b>DA.2</b> | <b>Programar l'algoritme de deep learning</b> | <b>70h</b>   | DA.1               | -               |
| DA.2.1      | Investigació                                  | 28h          | -                  | I               |
| DA.2.2      | Generació de l'algoritme                      | 42h          | DA.2.2             | P               |
| <b>DA.3</b> | <b>Avaluació de l'algoritme</b>               | <b>7h</b>    | DA.2               | -               |
| DA.3.1      | Definició de paràmetres                       | 5h           | -                  | A               |
| DA.3.2      | Avaluació                                     | 2h           | DA.3.1             | A               |
| <b>AA</b>   | <b>Aplicació de l'algoritme</b>               | <b>133h</b>  | <b>DA</b>          |                 |
| <b>AA.1</b> | <b>Generar un plugin de wordpress</b>         | <b>56h</b>   | -                  | -               |
| AA.1.1      | Investigació                                  | 28h          | -                  | I               |
| AA.1.2      | Generació del plugin                          | 28h          | AA.1.1             | P               |
| <b>AA.2</b> | <b>Aplicació en una plataforma</b>            | <b>70h</b>   | AA.1               | -               |

|             |   |             |            |   |
|-------------|---|-------------|------------|---|
| AA.2.1      | Investigació                                | 28h         | -          | I |
| AA.2.2      | Creació de la plataforma                    | 28h         | AA.2.1     | P |
| AA.2.3      | Aplicació del plugin a la plataforma        | 14h         | AA.2.3     | P |
| <b>AA.3</b> | <b>Avaluació del funcionament de l'eina</b> | 7h          | AA.1, AA.2 | - |
| AA.3.1      | Investigació dels criteris                  | 5h          | -          | A |
| AA.3.2      | Avaluació del funcionament                  | 2h          | AA.3.2     | A |
| -           | <b>TOTAL</b>                                | <b>504h</b> | -          | - |

Figura 2:

Figura 2: Taula resum de les tasques, les seves dependències i hores de dedicació estimades.

### 5.3 Diagrama de Gantt

El diagrama de Gantt ve creat segons les dependències especificades en el punt 5.1 on s'estableixen les dependències de forma concreta per a cada tasca. Com és pot veure, moltes de les tasques d'investigació no són dependents d'altres i es podrien realitzar de manera concurrent. Això no surt reflectit en el diagrama de Gantt doncs de cara a la realització del procés prioritzem el arribar a uns objectius per a poder considerar el treball com un èxit.

Amb això ens referim a que l'objectiu principal del projecte és el desenvolupament de l'algoritme i de cara a tindre temps d'afrontar imprevistos i dificultats, aquestes tasques prenen precedència a les altres tasques que compleixen objectius secundaris com és la aplicació del algoritme en una plataforma.

Com s'explicarà de manera extensa en el punt 5.4 sobre la gestió del risc, en cas de dificultats algunes de les tasques podrien fins i tot desaparèixer de cara a poder finalitzar el treball a temps i és per aquest motiu que dites tasques no es començarien a fer fins acabar les essencials, ja que podrien ser modificades o fins i tot eliminades per entrar en el termini esperat.

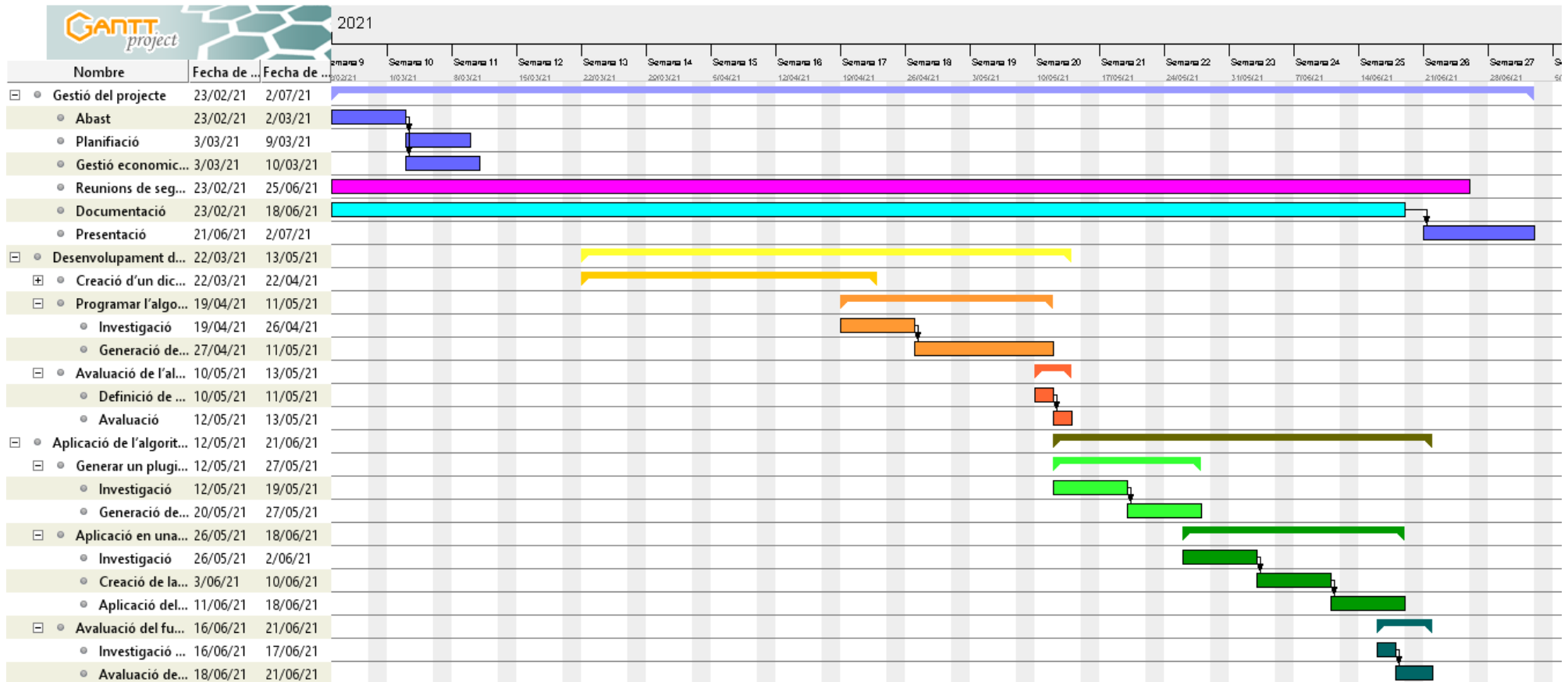


Figura 3: Diagrama de Gantt del projecte

## 5.4 Gestió del risc

Una part molt important de la planificació del temps és preveure els riscos i obstacles amb els que ens podem topar al llarg del desenvolupament del projecte, i planificar solucions i possibles plans alternatius a aquests per a assegurar un correcte i exitós projecte. L'idea és intentar complir el termini especificat, en cas d'impossibilitat, en les reunions de seguiment determinar si se li dona una mica més de temps per a finalitzar la tasca, o si per altra banda la distancia és insalvable i s'opta per un dels plans alternatius especificats a continuació:

1. *Dificultats en la generació del data set*

Tenint en compte el nivell de inexperiència, en el cas de que determinar un data set des de zero resultes una tasca massa àrdua, es podria usar data sets existents i orientats al objectius similars, tot i que aquests fossin en altres idiomes, per a adaptar-los i utilitzar-los com a base del nostre data set.

2. *Dificultats en l'etiquetatge del data set*

Per evitar grans endarreriments en les tasques d'etiquetatge, un dels plans alternatius per a solucionar, seria simplificar el nombre d'etiquetes amb els que definir els atributs del data set a dos: positiu o negatiu, de manera que la tasca d'etiquetatge fos molt més simple que determinar la puntuació d'un sentiment més específic.

3. *Dificultats en la creació d'una plataforma*

Les dificultats en les que ens podem trobar a la creació d'una plataforma on provar el plugin poden ser nombroses partint per la limitació del temps amb les que comptem, i per la desconexença i possible complexitat de la tasca.

Una solució amb la que comptem és usar una plataforma ja creada prèviament amb un companya i que consisteix en una plataforma similar a Twitter que permet la publicació d'estats i que podrien ser analitzats amb l'algoritme que desenvolupa aquest projecte. Aquesta seria una solució viable ja que també coneixeríem l'arquitectura de la plataforma, però que podria ser complexa ja que no ha estat dissenyada per aplicar aquest algoritme.

Els compliment del termini és un requisit imprescindible del treball de fi de grau. En cas de que totes les solucions alternatives no siguin suficients o el retard sigui insalvable, per a aconseguir finalitzar el projecte de manera satisfactòria és buscaria complir el seu principal objectiu. El objectiu imprescindible del projecte és el de crear un algoritme de processament del llenguatge natural que permeti determinar el sentiment dins d'un text, per entrar el termini i com a últim recurs, s'eliminarien totes les tasques d'aplicació de l'algoritme conformant el projecte únicament de la generació de l'algoritme i no demostrant les possibles aplicacions d'aquest.

## 5.5 Canvis a la planificació inicial

Com va ser examinat inicialment al generar la planificació temporal del projecte, es van valorar diferents riscos que podrien comportar un endarreriment substancial de l'avanç del treball. Per a intentar prevenir aquesta situació es va donar un marge d'una setmana en la finalització del projecte així com atribuir més hores de dedicació en les tasques que es consideraven de més risc. Finalment en el cas de que aquest endarreriment es produís i fos d'una magnitud suficientment gran per a que les distàncies resultessin insalvables, es resoldria a buscar la completació del principal objectiu del projecte, generar un algorisme de detecció de ciberbullying determinant la intencionalitat del text, i s'eliminaria de les tasques la part complementària d'aplicar aquest algorisme a una xarxa social actual per a poder veure una possible aplicació de l'algorisme generat en el món real.

Inicialment el projecte es dividit en dues parts: la primera i principal amb l'objectiu de generar un algorisme de detecció de ciberbullying en textos en català, i la segona: amb l'objectiu d'aplicar-ho a text en les xarxes socials.

Sobre aquesta primera part, era dividida principalment en tres blocs:

- Generació del data set
- Generació de l'algorisme
- Training i avaluació dels resultats

El projecte va avançar com era previst si més no amb alguna que altra dificultat inesperada, fins arribar al punt de training o avaluació dels resultats. Allà el percentatge d'encert de predicció de l'algorisme va ser molt inferior de l'acceptable, pel qual vam haver de realitzar diferents tasques no contemplades i costoses de temps com van ser:

- Revisió del data set
- Revisió de l'etiquetatge
- Revisió del model generat
- Expansió del data set
- Expansió de l'etiquetatge del data set
- Variació del model generat

L'objectiu d'aquestes tasques consistia en millorar l'alimentació del model o el model en sí, de manera que s'obtingués un bon resultat de predicció. Al no aconseguir-se es posa en contacte amb diferents persones incloent el tutor i companys d'aquest amb coneixements en el sector. Es va determinar aleshores que el motiu que el model no entreni es principalment el data set amb el que s'alimenta.

### 5.5.1 Problema

Es considera que el data set generat no és suficient correcte per a poder entrenar l'algorisme desenvolupat, cosa que ocasiona dos problemes: primerament, l'eina no funciona, i segon, endarrerix l'avanç a altres tasques i per tant, a la segona part del bloc, aplicació d'aquesta.

S'ha determinat que dins dels motius pel qual el data set no aconsegueix entrenar el model, es troben: el gran cost de temps per a generar un nombre molt petit de dades -hora i mitja per

traducció i puntuació de 100 oracions-, i la impossibilitat de donar-li una mida suficient gran dins les constriccions del temps del projecte, la subjectivitat del vocabulari i per tant imprecisió en el sistema de puntuació d'oracions, i per últim, poca especificació de les dades en el sector del cyberbullying i possible abast limitat de la variació del llenguatge en les oracions de data set.

### 5.5.2 Solució decidida

Finalment, i amb el consens del director del projecte, es decideix optar per a la solució plantejada en cas de risc del projecte on ens mantindrem treballant en aplicar diferents accions sobre el data set generat i d'optimització del processat per a buscar obtenir una millora en els resultats de predicció del model. D'entre aquestes optimitzacions que realitzarem es troben:

- Re càlcul del percentatge d'encert amb un percentatge més elevat de permissivitat entre els diferents etiquetes
- Simplificació del mètode d'etiquetatge (reduir les 20 etiquetes a 3 o 2) i juntament amb unes keywords, buscar la detecció autònoma dels tipus de text amenaçants.
- Canvi en les paraules i metodologia de distribució de les dades de data set per a fer training i testing buscant obtenir millors resultats.

Al decidir-nos per aquesta solució, els objectius principals de desenvolupament d'aquesta eina segueixen sent els mateixos, però es treu fora de l'àmbit del projecte l'aplicació d'aquest model generat en una xarxa social. Per tant, es mantenen els objectius però canvia la planificació, eliminant les tasques del segon bloc i afegint tasques d'optimització d'algorisme i anàlisi de resultat, deixant el llistat de tasques en el següent:

| <b>Id.</b>  | <b>Tasques</b>                                | <b>Temps</b> | <b>Dependència</b> | <b>Recursos</b> |
|-------------|---|--------------|--------------------|-----------------|
| <b>GP</b>   | <b>Gestió del projecte</b>                    | <b>180h</b>  | -                  |                 |
| GP.1        | Abast   | 28h          | -                  | CP              |
| GP.2        | Planificació                                  | 24h          | GP.1               | CP              |
| GP.3        | Gestió econòmica i de sostenibilitat          | 20h          | GP.1               | CP              |
| GP.4        | Documentació fase inicial                     | 5h           | GP.1,GP.2,GP.3     | CP              |
| GP.5        | Reunions de seguiment                         | 15h          | -                  | CP, I, P, A     |
| GP.6        | Documentació                                  | 60h          | -                  | CP, I, P, A     |
| GP.7        | Presentació                                   | 28h          | GP.4, GP.6         | CP              |
| <b>DA</b>   | <b>Desenvolupament de l'algorisme</b>         | <b>332h</b>  | <b>GP</b>          |                 |
| <b>DA.1</b> | <b>Creació d'un diccionari com a data set</b> | 136h         | -                  | -               |
| DA.1.1      | Investigació                                  | 32h          | -                  | I               |
| DA.1.2      | Elaboració d'un data set diccionari           | 56h          | DA.1.1             | P               |
| DA.1.3      | Mètode d'avaluació per l'etiquetatge          | 28h          | DA.1.2             | I               |
| DA.1.4      | Etiquetatge dels elements del data set        | 20h          | DA.1.3             | P               |
| <b>DA.2</b> | <b>Programar l'algorisme de deep learning</b> | 97h          | DA.1               | -               |

|             |                                      |             |        |      |
|-------------|--------------------------------------|-------------|--------|------|
| DA.2.1      | Investigació                         | 28h         | -      | I    |
| DA.2.2      | Generació de l'algoritme             | 42h         | DA.2.1 | P    |
| DA.2.3      | Investigació d'altres classificadors | 27h         | DA 2.2 | I,P  |
| <b>DA.3</b> | <b>Avaluació de l'algoritme</b>      | 17h         | DA.2   | -    |
| DA.3.1      | Definició de paràmetres              | 7h          | -      | I    |
| DA.3.2      | Entrenament dels models              | 6h          | DA3.1  | P    |
| DA.3.3      | Avaluació                            | 4h          | DA.3.2 | A    |
| <b>DA.4</b> | <b>Optimitzacions de resultats</b>   | 54h         | DA.3   | -    |
| DA.4.1      | Optimitzacions del data set          | 24h         | -      | A, P |
| DA.4.2      | Optimitzacions del model             | 16h         | DA.4.1 | A, P |
| DA.4.3      | Optimitzacions de la lemmatització   | 6h          | DA.4.2 | A, P |
| DA.4.2      | Optimitzacions de l'entrenament      | 8h          | DA.4.3 | A, P |
| <b>DA.5</b> | <b>Avaluació de resultats</b>        | 28h         | DA.4.4 | -    |
| DA.5.1      | Determinació de les mètriques        | 12h         | -      | I    |
| DA.5.2      | Avaluació dels nous resultats        | 16h         | DA.5.1 | A, P |
| -           | <b>TOTAL</b>                         | <b>512h</b> | -      | -    |

*Figura 4: Taula de llistat de tasques del projecte actualitzada*

Finalment, en quant als costos del projecte, aquest canvi en la planificació no ha suposat cap cost econòmic extraordinari per a l'autora, però sí ha suposat un increment de les hores de dedicació d'aquest. Es pot afirmar aleshores, que si aquest projecte es duagués a terme en una empresa amb empleats, suposaria l'augment del cost per l'augment d'hores de treball.

### 5.5.3 Diagrama de Gantt

A partir de les noves tasques i plans generats creem el nou diagrama de Gantt per a reflectir finalment quina ha sigut l'avanç temporal del projecte.

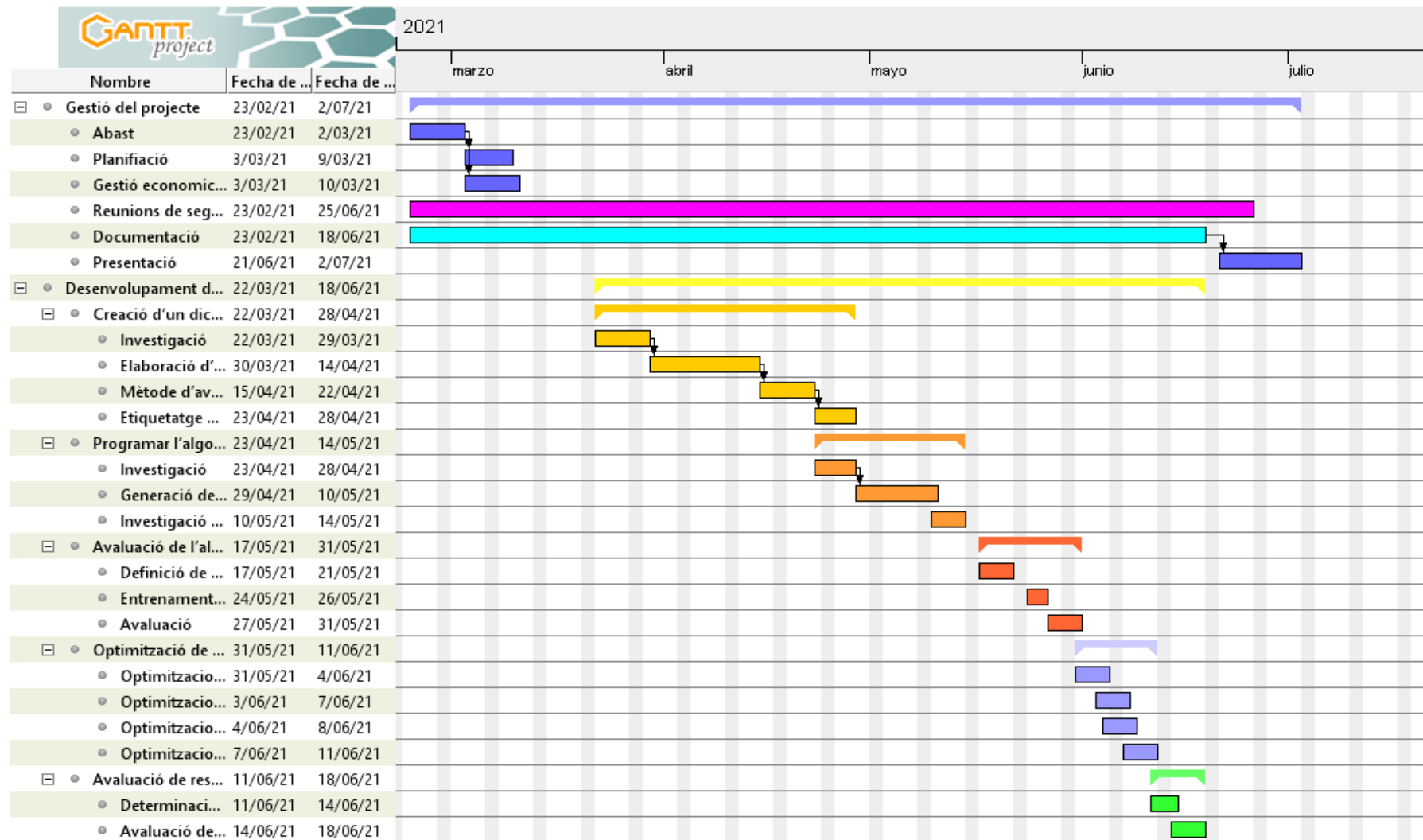


Figura 5: Diagrama de Gantt actualitzat

## 6. Gestió Econòmica

La gestió econòmica estima els costos necessaris per el desenvolupament del projecte una vegada feta la planificació temporal d'aquest. Amb la fi d'identificar aquests costos correctament hem de considerar els següents factors: recursos humans, hardware i software que es necessiten per a dur a terme el projecte, així com a una estimació de possibles costos generals i impostos. A aquests factors, per intentar generar un pressupost el més realista possible tenim en compte també les contingències i possibles imprevistos que poden sorgir. Per últim definim el control de gestió on especifiquem els mecanismes establerts per al control del pressupost generat.

### 6.1 Pressupost

#### 6.1.1 Recursos humans

Els costos de recursos humans fan referència als costos de salaris de treballadors. Aquests salaris venen determinats per els diferents rols que hi ha a la empresa. Aquest projecte actualment és dut a terme únicament per una persona, però per a realitzar el pressupost ens imaginarem un equip de quatre persones on cadascú té un dels quatre rols definits en la entrega anterior i que en realitat son tots duts a terme per l'autor del treball.

En els costos del treballador podem distingir diferents conceptes: El salari brut del treballador és la remuneració d'aquest abans d'aplicar-li els impostos corresponents com son l'IRPF o la seva contribució a la seguretat social. El sou que s'obté una vegada aplicats és l'anomenat salari net que és els diners que finalment li queden al empleat. El sou brut però no és el cost total de la empresa sinó que aquesta també ha de tributar a seguretat social un 33% del salari i per tant el cost per empleat es el salari brut x 1.33.

| Rol                  | Salari/hora (brut) | Salari/hora (net) | Salari brut + SS |
|----------------------|--------------------|-------------------|------------------|
| Cap de projecte (CP) | 27€/h              | 20.79€/h          | 35.91€/h         |
| Investigador (I)     | 21.5€/h            | 16.55€/h          | 28.59€/h         |
| Programador (P)      | 18.2€/h            | 14.01€/h          | 24.20€/h         |
| Avaluador (A)        | 18.2€/h            | 14.01€/h          | 24.20€/h         |

Figura 6: Taula de sous de creació pròpia basada en la guia salarial de Hays<sup>14</sup>

A continuació mostrem la taula amb les tasques determinades en la planificació temporal de manera que puguem determinar quin rol realitza cada tasca i el cost per tasca i el cost final en sou de treballadors. El cost que calculem és el cost de salari brut més seguretat social ja que és aquest el que pagarà la empresa i per tant el cost real estimat.

<sup>14</sup> MOTTO Digital Agency, S. (2021). Calculadora Salarial | Empresas: Guía del Mercado Laboral 2021 - Hays. Retrieved March 2021, from <https://guiasalarial.hays.es/empresa/calculadora-salarial/resultado>

| <b>Id.</b>  | <b>Tasques</b>                                | <b>Temps</b> | <b>Rol</b>  | <b>Cost + SS</b>  |
|-------------|---|--------------|-------------|-------------------|
| <b>GP</b>   | <b>Gestió del projecte</b>                    | <b>180h</b>  |             | <b>5887,36€</b>   |
| GP.1        | Abast   | 28h          | CP          | 1005,48€          |
| GP.2        | Planificació                                  | 24h          | CP          | 861,84€           |
| GP.3        | Gestió econòmica i de sostenibilitat          | 20h          | CP          | 718,20€           |
| GP.4        | Documentació fase inicial                     | 5h           | CP          | 179,55€           |
| GP.5        | Reunions de seguiment                         | 15h          | CP, I, P, A | 423,36€           |
| GP.6        | Documentació                                  | 60h          | CP, I, P, A | 1693,45€          |
| GP.7        | Presentació                                   | 28h          | CP          | 1005,48€          |
| <b>DA</b>   | <b>Desenvolupament de l'algoritme</b>         | <b>191h</b>  |             | <b>4990,96€</b>   |
| <b>DA.1</b> | <b>Creació d'un diccionari com a data set</b> | 114h         | -           | -                 |
| DA.1.1      | Investigació                                  | 28h          | I           | 800,52€           |
| DA.1.2      | Elaboració d'un data set diccionari           | 38h          | P           | 919,60€           |
| DA.1.3      | Mètode d'avaluació per l'etiquetatge          | 28h          | I           | 800,52€           |
| DA.1.4      | Etiquetatge dels elements del data set        | 20h          | P           | 484,00€           |
| <b>DA.2</b> | <b>Programar l'algoritme de deep learning</b> | 70h          | -           | -                 |
| DA.2.1      | Investigació                                  | 28h          | I           | 800,52€           |
| DA.2.2      | Generació de l'algoritme                      | 42h          | P           | 1016,40€          |
| <b>DA.3</b> | <b>Avaluació de l'algoritme</b>               | 7h           | -           | -                 |
| DA.3.1      | Definició de paràmetres                       | 5h           | A           | 121,00€           |
| DA.3.2      | Avaluació                                     | 2h           | A           | 48,40€            |
| <b>AA</b>   | <b>Aplicació de l'algoritme</b>               | <b>133h</b>  |             | <b>3464,44€</b>   |
| <b>AA.1</b> | <b>Generar un plugin de wordpress</b>         | 56h          | -           |                   |
| AA.1.1      | Investigació                                  | 28h          | I           | 800,52€           |
| AA.1.2      | Generació del plugin                          | 28h          | P           | 677,60€           |
| <b>AA.2</b> | <b>Aplicació en una plataforma</b>            | 70h          | -           | -                 |
| AA.2.1      | Investigació                                  | 28h          | I           | 800,52€           |
| AA.2.2      | Creació de la plataforma                      | 28h          | P           | 677,60€           |
| AA.2.3      | Aplicació del plugin a la plataforma          | 14h          | P           | 338,80€           |
| <b>AA.3</b> | <b>Avaluació del funcionament de l'eina</b>   | 7h           | -           | -                 |
| AA.3.1      | Investigació dels criteris                    | 5h           | A           | 121,00€           |
| AA.3.2      | Avaluació del funcionament                    | 2h           | A           | 48,40€            |
| -           | <b>TOTAL</b>                                  | <b>504h</b>  |             | <b>14.342,76€</b> |

*Figura 7: Taula amb els costos estimats de les tasques de la planificació temporal*

### 6.1.2 Hardware

Per a poder dur a terme el projecte necessitem de recursos hardware bàsics. En concret aquest hardware seria necessari per a cadascun dels membres del projecte (quatre si comptem a membre per rol) en aquest cas però tots els rols son duts a terme únicament per una persona.

Per a calcular el cost adient del hardware en els mesos de desenvolupament del projecte suposant una amortització de 4 anys com és permès per Hisenda hem de tindre en compte que en un any hi ha 220 dies laborables, el que comporta 1760 hores de treball anuals en una jornada de 8 hores de dedicació diària. Amb aquestes dades podem calcular la amortització com a :

$$\text{Amortització} = \frac{\text{Cost hardware } \text{€}}{4 \text{ anys} * 1760 \text{h laborables any}} * 504 \text{ h projecte}$$

A continuació mostrem la taula dels costos hardware suposats, la amortització no es calcula individualment ja que els costos inferior a 300€ es consideren baixos i van directament a costos directes. Es a dir, amortitzant un a un els aparells hardware, amortitzaríem el portàtil mentre que els altres elements serien costos directes, el procediment habitual és amortitzar sobre la suma doncs el cost si que és elevat.

| Hardware     | Cost        |
|--------------|-------------|
| Portàtil     | 800€        |
| Ratolí       | 7€          |
| Teclat       | 25€         |
| Monitor      | 100€        |
| <b>Total</b> | <b>932€</b> |

*Figura 8: Taula de costos estimats de hardware.*

Tenim aleshores un cost total de hardware de 932€, si calculem la amortització d'aquest obtenim un **cost final de 66.72€** .

### 6.1.3 Software

Els costos del projecte en quant al software necessari usat és ínfim ja que el projecte es desenvolupa majoritàriament sobre software lliure el qual es gratuït. Per tant els únics costos de software que tindrem seran:

| Software     | Cost        |
|--------------|-------------|
| Windows      | 90€         |
| Office 365   | 67€ anuals  |
| <b>Total</b> | <b>157€</b> |

*Figura 9: Taula de costos software estimats*

Tot i que hisenda permet amortitzar el software a dos anys, aquests costos no seran amortitzats ja que com hem mencionat prèviament i segons el pla general comptable (PGC) no s'amortitzen valors inferiors a 300€.

#### 6.1.4 Costos generals

Per a finalitzar amb els costos que podríem trobar quan generem un projecte, tenim els anomenats costos generals. Aquests inclouen diferents tipus de cost com son els gestos d'internet, aigua, electricitat, i material d'oficina. Aquest es un projecte que no necessita d'estructura, per tant idealment aquest treball es duria a terme a un *coworking*. Aquests son espais compartits on professionals independents desenvolupen el seu treball i son altament comuns a la actualitat doncs dins d'una tarida general t'inclouen tots aquests costos generals (a excepció del material d'oficina). La tarifa d'un espai coworking varia segons factors com la localització d'aquests però es mantenen sovint en una tarifa de 300€ al mes. Tenint en compte que el projecte es durà a terme al llarg de 5 mesos, el cost total del espai serà de 1500€. A aquest cost se li poden afegir costos variables com el material d'oficina que serien uns possibles 15€ al mes.

| <b>Costos generals</b> | <b>Cost</b>  |
|------------------------|--------------|
| Coworking              | 1500€        |
| Material d'oficina     | 75€          |
| <b>TOTAL</b>           | <b>1575€</b> |

*Figura 10: Taula dels costos generals estimats*

#### 6.1.5 Contingència

El cost de contingència és el sobre cost calculat per a cobrir les despeses generades per obstacles i imprevistos. La probabilitat de trobar problemes en un treball d'investigació amb un termini fixe és elevat. Hem decidit fixar aleshores un cost de contingència del 15% per a fer front a problemes durant el desenvolupament del projecte.

| <b>Típus</b>    | <b>Cost</b>       | <b>Contingència €</b> |
|-----------------|-------------------|-----------------------|
| Recursos humans | 14.342,76€        | 2151,41€              |
| Hardware        | 932€              | 139,80€               |
| Software        | 157€              | 23,55€                |
| Costos generals | 1575€             | 236,25€               |
| <b>TOTAL</b>    | <b>17.006,76€</b> | <b>2551,01€</b>       |

*Figura 11: Taula dels costos estimats amb un 15% de contingència*

#### 6.1.6 Imprevistos

L'última part dels costos son els imprevistos. Aquests tracten els costos que poden sorgir durant el desenvolupament del projecte per conseqüència dels riscos i obstacles del desenvolupament prèviament analitzats. Aquests principalment son:

- *Augment del temps de desenvolupament:*  
Aquest és un dels imprevistos amb més possibilitat de succeir ja que nombrosos obstacles com la inexperiència i el desconeixement ens porten a aquesta conseqüència. Estimem aleshores un 25% de probabilitats de que succeeixi. A més, es calcula que com a màxim augmentaríem el tems de desenvolupament en dues setmanes abans de començar a retallar el abast. Això ens porta a un cost de 48 hores de programador el que comporta un augment de 1161,6€.
- *Error de hardware:* En cas de que s'espalli algun element del hardware necessari per a realitzar el desenvolupament del treball aquest s'haurà de reemplaçar. LA probabilitat però de que això succeeixi es considera mínima (estimem un 5%) ja que els dispositius es tracten com a nous.

| <b>Imprevist</b>   | <b>Cost</b>    | <b>Risc</b> | <b>Cost total</b> |
|--------------------|----------------|-------------|-------------------|
| Augment temporal   | 1161,6€        | 25%         | 290,4€            |
| Ordinador Portàtil | 800€           | 5%          | 40€               |
| Ratolí             | 7€             | 5%          | 0,35€             |
| Teclat             | 25€            | 5%          | 1,25€             |
| Monitor            | 100€           | 5%          | 5€                |
| <b>TOTAL</b>       | <b>2093,6€</b> | -           | <b>337€</b>       |

*Figura 12: Taula del cost afegit per imprevistos*

### 6.1.7 Cost total

Una vegada vists tots els costos del projecte, la suma d'aquests correspon a el cost total que tenim, que suma 19.894,77euros.

| <b>Tipo</b>     | <b>Cost</b>       |
|-----------------|-------------------|
| Recursos humans | 14.342,76€        |
| Hardware        | 932€              |
| Software        | 157€              |
| Costs generals  | 1575€             |
| Contingència    | 2551,01€          |
| Imprevistos     | 337€              |
| <b>TOTAL</b>    | <b>19.894,77€</b> |

*Figura 13: Taula de pressupost final estimat del projecte*

L'objectiu és que el projecte generi suficients ingressos per a generar beneficis. Part d'aquests anirien dirigits a la reinversió per el manteniment i seguir desenvolupant el projecte. De mateixa manera els sobre costos calculats que no s'hagin gastat seran altra vegada invertits com a capital de la empresa.

## 6.2 Control de gestió

És necessari definir els indicadors i mecanismes de control que permetran detectar i analitzar si durant el desenvolupament del projecte aquest s'até al pressupost ideat. Amb aquest objectiu, durant tot el treball anirem mantenint un registre de les hores dedicades, similars a la obligació de fitxar en el treball, per a tindre un control real de les hores finalment dedicades i apreciar si ens estem mantenint dins el nostre pressupost. Per a fer-ho definim les següents mètriques:

- *Desviació de la realització de tasques:*  
 $(\text{hores estimades} - \text{hores reals}) * \text{cost estimat}$
- *Desviació del cost de la realització de tasques:*  
 $\text{cost estimat tasques} - \text{cost real tasques}$
- *Desviació de costos en recursos:*  
 $\text{cost estimat recursos} - \text{cost real recursos}$
- *Desviació total del cost d'imprevistos:*  
 $\text{cost estimat imprevistos} - \text{cost real imprevistos}$
- *Desviació total d'hores:*  
 $\text{hores estimades} - \text{hores reals}$

La mètrica que serà més usada durant el desenvolupament del projecte serà la de la desviació de les hores. Al llarg del desenvolupament del projecte quan finalitzem una tasca, comprovarem en la planificació temporal que aquesta s'hagi complert en el temps estimat. En el cas que això no succeeixi ens trobem en dues situacions: la tasca s'ha completat amb menor temps del esperat, cosa que serà una desviació positiva del projecte i no requerirà d'accions. O en el cas contrari, si la tasca hagi tingut una durada superior a la estimada, s'hauran de valorar si aquesta desviació negativa es suficientment important com per a prendre decisions que alterin substancialment la planificació del projecte o, si en cas contrari, es pot seguir endavant amb mínims canvis. Aquesta mètrica serà aleshores essencial durant tot el desenvolupament del treball per a poder assegurar una entrega del projecte en el termini indicat.

## 7. Sostenibilitat i compromís social

En qualsevol projecte d'avui dia és necessari analitzar la sostenibilitat del projecte tenint en compte les tres principals dimensions que conformen aquesta: econòmica, ambiental i social. En aquest apartat és realitza una autoavaluació dels coneixements que té l'autor sobre la competència de sostenibilitat i a continuació s'analitzen els tres àmbits dins del marc d'aquest projecte en concret.

## 7.1 Autoavaluació

El concepte de sostenibilitat ens ha anat seguint al llarg de tot el grau d'Informàtica sense fer-hi massa èmfasi però sempre tenint-lo present. Sobre el àmbit de la sostenibilitat ambiental en soc conscient dels cementiris de residus tecnològics existents, del gran problema que és la obsolescència programada i de la necessitat de donar una llarga vida als productes i en cas de llençar-los, reciclar-los adequadament per a poder donar-los una nova vida. Aquesta lliçó va quedar bastant gravada en una activitat al segon curs de la carrera on es netejaven dispositius antics per a poder reutilitzar-los.

Sobre l'apartat social, potser no soc tant conscient d'haver-lo treballat al llarg del grau, però considero que es un sector de la sostenibilitat que s'entén amb més facilitat i a la que la gent és més propensa. Considero que és un impuls natural el voler crear i desenvolupar projectes que siguin útils en un àmbit social i que millorin la vida de les persones i que tot i que no tinguem els coneixements per a fer un anàlisis tècnic de la sostenibilitat social d'un projecte, el concepte i idea general és fàcil d'aconseguir.

Finalment, sobre la dimensió econòmica de la sostenibilitat considero que pot ser la dimensió que més controlo tècnicament, degut als meus estudis previs d'economia al cursar el batxillerat social. Els conceptes d'amortització, costos, assentaments comptables, impostos, DAFOs.. són conceptes amb els que he treballat prèviament i amb els que estic còmoda.

Fent una visió global de la sostenibilitat però considero que tot i que tinc una idea general dels tres grans àmbits que la comprenen, em falten les capacitats per a poder fer un anàlisis tècnic i objectiu de la sostenibilitat de un projecte comprnent les seves tres dimensions. Tot i que penso que podria analitzar bastant bé la viabilitat econòmica d'un projecte, la sostenibilitat d'aquest ve donada per les tres dimensions les quals estan fortament lligades i la meva manca de coneixença en aquestes em faria inviable fer un anàlisis amb condicions. Per altra banda, ja el tindre uns coneixements base de les tres dimensions es pot considerar un èxit ja que és un punt de partida sobre la qual investigar i poder aprendre per el meu propi compte.

## 7.2 Estudi de l'impacte econòmic

Una part molt important a l'hora de fer una proposta de projecte per a que aquesta sigui aprovada és el cost econòmic del treball. Si aquest no té un cost viable, ja sigui perquè és massa elevat o perquè surt de les capacitats de la empresa, el projecte no es podrà dur a terme. En el nostre cas, el projecte es podria desenvolupar amb un grup reduït de gent però l'anàlisi del cost seguiria sent necessari per a assegurar la viabilitat d'aquest i la possibilitat d'obtenir beneficis i per tant que fos un projecte viable.

Per a estimar els costos s'ha usat un sou mitjà elevat, que seria l'ideal si el projecte es desenvolupés en una empresa de forma real, però actualment s'ha dut a terme únicament per una persona que s'encarrega de tot i com a treball de final de grau cosa que comporta que no hi hagi una dedicació absoluta al projecte i per tant les hores de dedicació i el grau de coneixement i especialització en el entorn i de les eines que s'usen no és el més ideal i que potser

es buscaria en un entorn de contractació real per evitar els majors riscos que eren la desconeixença i falta d'experiència.

Com ha estat comentat anteriorment, la idea del projecte explota el que es considera un forat en el mercat. En la societat actual, a Catalunya hi ha un problema que és el ciberbullying a nens, i no s'ha generat una manera fàcil i adient per a solucionar-ho. L'objectiu d'aquest projecte es aconseguir desenvolupar una solució viable per a eradicar aquest tipus de solucions a Catalunya. Econòmicament al considerar que aprofita un forat en el mercat existent, ja que no existeixen eines que compleixen el objectiu en l'entorn català, seria viable ja que no hi ha competència amb altres solucions existents prèvies, el que ens permetria suposar beneficis al no haver-hi alternatives al mercat, i permetre ser una mica laxos amb els costos ja que no hi hauria necessitat de preus altament competitius com seria el cas d'un producte en un mercat sobre explotat.

A més ha estat pensat i dut a terme en aproximadament cinc mesos, el que es un temps bastant limitat, sobretot tenint en compte que només hi ha una persona desenvolupant el projecte. Aquest es podria realitzar en menys temps augmentant el nombre de persones que desenvolupessin el treball però augmentaria el cost d'aquest. Per altra banda la disminució dels costos només es podria donar amb la reducció del sou als empleats (cosa negativa ja que els sous estan estimats a preu de mercat actual) o eliminant el rol d'investigador i passant aquesta feina al programador, el que es una mala praxis ja que no es correcte donar feina a un empleat que no li correspon per a retallar costos. Sobre els costos d'infraestructura aquests ja estan minimitzats al màxim, el hardware usat és el mínim necessari i el software empleat és en la gran major part software lliure i per tant sense cost. Sobre els costos d'estructura, aquests consisteixen en el cost del coworking, el qual es una alternativa molt més econòmica que el lloguer d'una oficina o altres alternatives.

Degut a la labor social a la que esta orientada la idea del projecte, idealment l'objectiu d'aquest no seria el benefici econòmic sinó tenir un cost representatiu per a poder recuperar la inversió de capital feta per el desenvolupament del projecte i usar la resta d'ingressos per a mantenir l'eina desenvolupada o afegir millores i adaptar-la si fes falta. S'estima que el cost del projecte per tant serà principalment en el desenvolupament d'aquest, ja que una vegada entrenat el model els costos computacionals no seran extremadament grans. D'aquesta manera no es considera que es pugui fer més viable ja que els algorismes de deep learning busquen ja la eficiència computacional. Afegir que la viabilitat del projecte no es probable que es vegi compromesa ja que busca un enfoc innovador i per tant en desenvolupament, cosa que comporta que aquest tipus de solucions no quedin obsoletes en un futur proper.

### 7.3 Estudi de l'impacte econòmic

El desenvolupament del projecte per compte pròpia i amb la independència que ve donada del desenvolupament d'un treball de fi de grau, ha donat experiència i realisme en el desenvolupament d'un projecte d'inici a fi aprenent totes les fases que comprenen un treball i

obligant-me a buscar recursos i desenvolupar la meua creativitat per a finalitzar-lo de manera satisfactòria, el que ha comportat un gran desenvolupament acadèmic personal. Per altra banda, degut a la naturalesa del projecte que busca fer una labor social com es combatre el ciberbullying, s'obté la satisfacció personal de si més no enfocar les energies en un projecte que enriqueix la vida de persones.

Actualment no existeix una eina que permeti comprendre i detectar els missatges en català i que per tant doni una solució aplicable aquí a Catalunya. La meua eina mantindria elements similars a les ja existents però seria generada en un idioma en la que eines com aquesta encara no han sigut creades. També busca fer un pas més endavant i no convertir-la en una eina de benefici propi sinó un software lliure que es pugui aplicar de manera generalitzada a tot tipus de contingut d'internet, ajudant no només a que les víctimes de ciberbullying es puguin defensar sinó que els propietaris de xarxes socials i altres plataformes puguin usar-la per a prendre mesures per a evitar aquestes situacions abans de que succeeixin.

Idealment, el meu projecte millorarà socialment la vida dels nens que es troben en risc o patint atacs de ciberbullying i serà un pas endavant en la eradicació d'aquestes practiques i en fer les xarxes socials un espai més segur per a tothom.

Com he mencionat anteriorment, el projecte va orientat a combatre el ciberbullying en menors d'edat. El ciberbullying és actualment un gran problema a Catalunya avui en dia i que pot causa danys irreparables a joves que moltes vegades es troben indefensos davant d'aquest problema. Hi ha una gran necessitat real d'eines com aquestes ja que el perill d'aquests atacs es cada vegada major i encara no hi ha s'ha donat una solució per a combatre'ls.

Personalment el treball ha comportat un repte personal bastant gran al endinsar-se en unes matèries en les quals no estava familiaritzada, i amb el repte de tenir una limitació temporal tant marcada. Aquestes dificultats si més no eren suavitzades amb la coneixença que el treball no només busca superar un repte acadèmic sinó que busca beneficiar substancialment la vida de les diferents persones en especial nens en situació de risc d'assetjament online.

Al ser un temari tant específic com és la identificació de ciberbullying, és difícil que es pugui capgirar el funcionament del projecte que busca la protecció i defensa de l'usuari per a dedicar-se a motius menys socials.

#### 7.4 Estudi de l'impacte ambiental

El projecte ja de per si ha estat pensat de manera que es pogués dur a terme amb la menor necessitat de material possible. Certament l'enfoc d'aquesta decisió no era buscar la reducció del impacte ambiental sinó la reducció dels costos del projecte però tot i no estar pensat en aquesta direcció les conseqüències d'aquest plantejament han significat un impacte ambiental ínfim. Tot i que en l'anàlisi pressupostari es comptabilitzen els costos de hardware, aquests

poden ser reutilitzats assumint que el treballador ja disposa del seu propi portàtil, ratolí, teclat i monitor, el que no seria estrany ja que son recursos bàsics i que es pot assumir que un treballador d'aquesta indústria segurament tindria. Així doncs com aquests son els únics recursos usats, es pot assumir que l'impacte ambiental ja és el mínim possible i que aquest serà molt baix.

Actualment les altres solucions existents que poden ser similars a les meves son també digitals. Tot i que no es possible saber els detalls del desenvolupament dels altres projectes, es pot assumir que el impacte mediambiental que aquests tenen no serà molt elevat ja que no son projectes de grans indústries si no més aviat petits. Per tant, desgraciadament no considero que ambientalment el meu treball millori les solucions existents ni tingui un impacte ambiental positiu, però si que considero que no és pitjor que les solucions que ja es troben i tampoc tenen un impacte ambiental negatiu.

### 7.5 Matriu de sostenibilitat

A continuació representem la matriu de sostenibilitat generada amb la puntuació considerada una vegada fet l'anàlisi que ha generat cada informe.

|                     | PPP                | Vida útil         | Riscs            |
|---------------------|--------------------|-------------------|------------------|
| Ambiental           | Consum del disseny | Petjada ecològica | Riscs ambientals |
|                     | 8                  | 17                | 0                |
| Econòmic            | Factura            | Pla de viabilitat | Riscs econòmics  |
|                     | 8                  | 18                | -3               |
| Social              | Impacte personal   | Impacte social    | Riscs socials    |
|                     | 10                 | 18                | 0                |
| Rang Sostenibilitat | 26                 | 53                | -3               |
|                     | 76                 |                   |                  |

Figura 14: Matriu de sostenibilitat

## 8. Marc legal

Aquest projecte esta desenvolupat com a Treball de Fi de Grau de la Facultat d'Informàtica de Barcelona i queda per tant subjecte a la normativa del projecte publicada, on s'especifica que la propietat intel·lectual i industrial d'aquests tipus de TFG ve regulada per la normativa aprovada pel Consell de Govern el (10/10/2008). En concret s'indica que:

“..s’aprova la confidencialitat, responsabilitat patrimonial i propietat industrial i intel·lectual a la Universitat Politècnica de Catalunya. D’aquesta normativa destaquem els paràgrafs següents relatius a les invencions i les obres dels estudiants dirigides o coordinades pel professorat de la UPC:

- ... correspondrà a la UPC la titularitat sobre les invencions desenvolupades exclusivament pels estudiants si s’ha desenvolupat en el marc d’una activitat acadèmica que hagi estat dirigida i/o coordinada pel professorat de la UPC.
- En el cas que el desenvolupament de l’obra intel·lectual hagi estat dirigida i/o coordinada pel professorat de la UPC, correspondrà a la UPC la titularitat dels drets d’explotació sobre aquesta obra i l’estudiant i el professor seran considerats coautors de la mateixa.
- En cas d’explotació de l’obra per part de la UPC que li suposi un benefici econòmic, l’autor o conjunt d’autors tindran dret a una participació del 50% dels beneficis nets obtinguts.

Mencionar també l’obligatorietat de l’ús ètic i amb compliment de les normatives de les llicències de les llibreries i productes open source que s’utilitzin en el desenvolupament d’aquest com qualsevol projecte.

## 9. Disseny de la solució

De cara al desenvolupament del projecte, cal prendre un seguit de decisions que determinaran el nostre avenç en el treball. En particular parlem del tipus de data set que es decidirà usar com també la mida i característiques d’aquest, i els classificadors amb els que treballarem així com decidir si volem implementar el nostre propi.

Inicialment després de la investigació pertinent sobre els recursos disponibles i sent conscients de la nostra desconeixença del sector i limitació temporal del treball, determinem que la mida del nostre data set estarà al entorn de les 7000 oracions les quals seran de diversa longitud. També es decideix provar un enfoc mixta de cara a la elaboració del classificador, intentant generar el nostre propi degut al valor afegit d’aprendre i conèixer completament l’interior del classificador amb el que es treballa però també provar i valorar altres classificadors existents de cara a comparar rendiments i conèixer d’altres alternatives ja implementades.

### 9.1 Anàlisi de les eines

A continuació es determinen el conjunt d’eines, llibreries i llenguatges usats per a la implementació del projecte partint de que aquest es du a terme en un ordinador portàtil amb el sistema operatiu de Windows.

Es decideix com a llenguatge per a la implementació usar Python. Tot i que l’autora no ha treballat mai amb aquest, no resulta un principal inconvenient ja que la sintaxis és relativament senzilla d’aprendre , a més Python compta amb la possibilitat d’usar una gran quantitat de

llibreries enfocades al desenvolupament de Machine learning i deep learning que es el que sentència el seu ús com a llenguatge de desenvolupament de la implementació del projecte.

Una vegada decidit el llenguatge de desenvolupament, decidim l'entorn d'execució. Aquest consistirà en Google Colab, el qual et proporciona un entorn de Jupyter Notebook sense requerir cap tipus de configuració i és executat al cloud, el format de creació del programa consisteix en l'anomenat quadern en format .ipynb que permet executar tant quaderns d'IPython, Jupyter i Colab, és a dir permet la execució de quaderns creats en els principals entorns d'execució de Python. Al executar el quadern ens connectem a un entorn proporcionat per una VM de Google Compute Engine que correspon a les màquines virtuals de Google en el cloud. A més, aquest entorn et permet escollir si vols crear els teus quaderns en Python 3 o Python2, en el nostre cas s'ha fet amb Python3 que és la configuració per defecte. A més, un avantatge principal és que et permet executar usant la GPU. Aquest ús de la GPU es troba limitat a un cert pes d'execució, si és molt prolongada i potent es pot desconnectar, però per la execució que duem a terme és més que suficient.

| EINES DEL PROJECTE      |   |
|-------------------------|---|
| <b>Hardware</b>         | Portàtil HP ENVY Notebook (Intell® Core (TM) i5-6200U CPU @ 2.30GHz)  |
| <b>Sistema Operatiu</b> | Microsoft Windows 10 Home   |
| <b>Software</b>         | Google Colab  |
| <b>Llibreries:</b>      | <ul style="list-style-type: none"><li>• Nltk</li><li>• Snowball</li><li>• Tensorflow</li><li>• Skit-Learn</li><li>• Numpy</li><li>• Pandas</li><li>• Matplotlib</li></ul> |

*Figura 15: Taula d'eines usades en el projecte*

Tot seguit expliquem les llibreries usades en la implementació del projecte i per a quines funcionalitats. Aquestes també seran especificades en el desenvolupament d'aquests.

### 9.1.1 Llibreria NLTK

Consisteix en un kit enfocat a oferir un conjunt de llibreries i programes pel processat del llenguatge natural amb l'objectiu de donar suport i ajudar a evolucionar aquest sector. Esta enfocat en el desenvolupament i creació de programes python per a aplicar deep learning als texts i ofereix eines que cobreixen tot el tractament de text en diferents llenguatges.

Proporciona una gran varietat de dades lingüístiques com llistats de *stopwords*, corpus, diferents models etc...<sup>15</sup> però malauradament no una gran quantitat de serveis estan en català.

De la llibreria nltk usem les seves funcionalitats ja implementades de:

- `Word_tokenize`: per a realitzar les accions de tokenització a nivell de paraula d'una oració.
- `Punkt`: per a descarregar les dades de puntuació de les oracions.
- `Stem`: des de la llibreria nltk també podem accedir al `SnowballStemmer` per les tasques de stemmatització.

### 9.1.2 Snowball

Consisteix en un llenguatge de processat de strings curts que busca crear diferents algorismes per a realitzar tasques de stemmatització i que conté una sèrie d'algorismes amb aquest objectiu ja implementats. Snowball esta actualment mantingut com un projecte de comunitat i cobreix un algorisme de stemmatització en Català aportat per una de les participants en el projecte. Aquest algorisme segons la documentació aportada comprova els seus bons resultats amb una llibreria de referència també en aquest sector que és `Freeling`. Per tant de cara el projecte usarem Snowball per a les tasques de stemming del nostre data set generat, en el preprocessat d'aquest.

### 9.1.3 Llibreria Tensorflow

És una llibreria de opencode que t'ofereix de diverses llibreries i eines per a fàcilment implementar un algorisme de deep learning. Aquesta serà usada per la creació del nostre propi classificador com una xarxa neuronal. Això ho fem utilitzant Tensorflow juntament amb Keras el que ens permet la creació del nostre model amb simples comandes per a generar el nostre model seqüencial amb les diverses capes denses, d'embedding i del tipus que necessitem per fer-lo. A més ens ofereix les diferents funcions de loss i optimització així com la possibilitat de fer el training dels nostres models i el seguiment de diferents mètriques com és la accuracy al llarg de l'entrenament.

D'entre les funcionalitats que usem es troben:

- `Keras.Sequential`: per la creació d'un model seqüencial
- `Keras.Layers`: per al creació dels diferents tipus de capes dels models
- `Callbacks`: per a aplicar les diferents mètriques i funcions al model
- `Compile, Fit, Train`: Per a generar el model, alimentar-lo amb les dades i entrenar-lo

### 9.1.4 Llibreria Scikit-Learn

És una altra de les llibreries open source enfocades en accions per al sector de Machine learning de Python i permet la generació de mode molt senzilla de diversos algorismes d'aprenentatge

---

<sup>15</sup> Per veure el llistat complet de serveis que nltk ofereix, referir-se a [http://www.nltk.org/nltk\\_data/](http://www.nltk.org/nltk_data/)

tant supervisat com autònom en Python. Aquesta llibreria esta constituïda i per tant inclou altres llibreries útils que usem com son Numpy, Pandas y Matplotlib. En el nostre projecte en concret ens basem en aquesta llibreria per l'ús de diferents algorismes classificadors i les mètriques d'aquests per a avaluar-los.

De les funcionalitats principals que usem, i de forma més específica trobem:

- `Train_test_split`: per a generar la divisió del data set entre la part d'entrenament i la de testing.
- `Feature_extraction`: Per a poder generar la matriu amb la funció `Count Vectorizer`
- `Model_Selection`: Per a trobar els paràmetres amb millor rendiment en l'entrenament d'un classificador.
- Importació dels cinc classificadors diferents a usar.
- `Metrics`: Per a poder obtenir les dades per a avaluar els diferents classificadors

### 9.1.5 Llibreria Numpy i Pandas

Aquestes son dos de les llibreries essencials per a treballar amb Python. Numpy consisteix en el tracte de les dades com si fossin matrius anomenades arrays, les quals son emmagatzemades en memòria de forma extremadament eficient. Pandas en canvi, es basa en la pròpia llibreria Numpy per a crear les estructures amb les que treballa les quals son series per una dimensió, DataFrames per estructures amb dues dimensions i Panel per a estructures amb tres dimensions. A aquestes estructures se li afegeixen noves funcionalitats que permeten el tractament fàcil i àgil de les dades amb les que es treballen.

### 9.1.6 Llibreria Matplotlib

És una de les principals llibreries que treballen amb Python i esta orientada a la creació de gràfics en 2D. Aquest és per tant l'ús que li donem ja que és usada per a representar gràficament les diverses mètriques que obtenim durant la implementació de diferents classificadors i en la avaluació d'aquests.

## 10. Implementació

Actualment, per a produir un algoritme de processat de sentiment en text en anglès es relativament senzill doncs s'han desenvolupat gran quantitat de llibreries i generat un enorme munt de datasets disponibles i ja catalogats amb polaritat que permeten fer aquest anàlisi de manera senzilla. Això però, no es dona en el català, doncs és un idioma que no es recolzat per gran majoria de les llibreries de processat de llenguatge natural com es nltk i que, en cas contrari, si es troba recolzat no està fet amb gran qualitat i quantitat de dades, de manera que no es pot assegurar un bon resultat en el processat de text en català amb les llibreries existents. Aquesta ha estat una de les dificultats principals del projecte, doncs estem realitzant el treball en un

idioma del qual no s'han generat ni produït gran majoria de dades i eines per a realitzar aquest processat del llenguatge natural.

Podem dividir el desenvolupament de l'algoritme en tres parts principals: la obtenció del data set usat per entrenar el nostre model, el preprocessat de les dades per a poder fer-les servir per l'entrenament, i la generació del model que s'encarrega de predir els resultats, en el nostre cas, si el text analitzat es positiu o negatiu.

A continuació s'expliquen els processos duts a terme i les especificacions, així com es justifiquen les decisions preses per a decidir aquestes especificacions i s'exposen en contraposició a les eines existents per a altres idiomes.

## 10.1 Dataset:

El dataset consisteix en el conjunt de dades usades com a base per a entrenar un algoritme amb l'objectiu que la màquina pugui prendre decisions autònomes sobre dades que no han sigut prèviament registrades.

Actualment no existeix cap corpus públic, estandarditzat o usat per a ninguna institució, de data sets en català, fet que ens obliga a generar el nostre propi data set. Per aquest motiu sorgeixen un conjunt de problemes:

1. La gran mida de data set que es requereix per a fer un bon entrenament de deep learning i el temps limitat del projecte. Es calcula de mitjana que es necessiten un conjunt de 7000 dades com a mínim.
2. La generació de la polaritat de les dades del data set. És necessari donar una puntuació entre -1 com a text molt negatiu i agressiu i 1 com a text positiu. Aquesta polaritat ha de ser determinada per uns criteris objectius i es de gran dificultat determinar-la manualment i de gran cost temporal per la gran quantitat de dades del data set.
3. La manca de llibreries per a determinar la polaritat d'un text en català i la poca efectivitat i fiabilitat de les eines usades que si recolzen l'idioma.
4. La varietat del data set i la rellevància de les dades usades. Les oracions han d'estar enfocades a determinar cyberbullying, i és preferible la major varietat d'estructures morfològiques possibles, de manera que el programa no aprengui de manera esbiaixada.

### 10.1.1 Opcions plantejades

Tenint en compte que no hi havia cap data set estandarditzat per a dur a terme un entrenament d'aquest tipus en català, i la nostra desconexió en l'àmbit, es van realitzar diferents cerques i opcions per a poder generar un data set correcte i funcional en el temps limitat que teníem. Entre les opcions que vam plantejar es troben:

-Us d'un data set de paraules catalanes ja puntuades: descartat, doncs no volem que l'algorisme aprengui a nivell de paraula sinó a nivell d'oració o fins i tot text.

- Us d'un data set de TASS: Aquesta opció consistia en agafar un dels datasets de TASS un taller d'anàlisi semàntic en la SEPLN <sup>16</sup>. Aquests datasets estan formats per un conjunt molt gran de tweets recopilats en castellà o en anglès i etiquetats amb 3 o 5 nivells de polaritat que consisteixen en: Negatiu, Neutre, Positiu, Molt negatiu o Molt Positiu. Al ser dades de tweets, la gent utilitza un llenguatge col·loquial, cosa que fa que les dades siguin realistes però a la vegada dificulta i empitjora considerablement la efectivitat de les traduccions automàtiques. Finalment es descarta, doncs les dades no estan centralitzades en la detecció de ciberbullying i agressivitat i les etiquetes ja aportades tampoc són útils pel mateix motiu. Un clar exemple seria la oració "No m'agrada la verdura" la qual no està assetjant a ningú però és una oració negativa.

### 10.1.2 Opció escollida

Finalment, s'ha decidit usar un data set <sup>17</sup>trobat online a la pàgina de Mandalay Data, un repositori que permet l'emmagatzematge i compartició de data. Aquest data set tenia com a objectiu l'entrenament d'algorismes enfocats a la detecció de ciberbullying i assetjament, i cobria una gran varietat de tipus d'agressió des de: toxicitat fins a sexisme. En particular, era un data set en anglès també obtingut com a resultat d'una recopilació de tweets i amb una polaritat que determinava si era assetjament o no. Aquesta polaritat, com venia explicat a la especificació del data set, havia estat determinada per un grup de tres persones. Cadascú individualment havia determinat si considerava que la oració en si era ofensiva, insultant o agressiva, i per tant si la considerava una frase de ciberbullying. Una vegada els tres membres havien donat la seva valoració es determinava per majoria si la oració era o no ciberbullying.

Per a usar aquest data set però, primer l'hem hagut de traduir al català. Això s'ha fet manualment ja que com s'ha explicat anteriorment, les traduccions automàtiques realitzen molts errors, sobretot en aquest tipus de dades recopilades de xarxes socials on els usuaris escriuen incorrectament, escurçant les paraules i utilitzant varietat de col·loquialismes i insults els quals no són traduïts correctament de manera automàtica. Un exemple d'això seria la paraula "bitch" la qual és traduïda de manera automàtica com a "gossa" en català, paraula que no té la mateixa connotació ofensiva i insultant que a l'anglès. A mesura que hem anat duent a terme aquestes traduccions, hem aprofitat per a revisar l'etiquetatge de positiu o negatiu i determinar si corresponia amb l'etiquetatge que buscàvem.

### 10.1.3 Etiquetatge del data set

Com es pot anar intuïnt, segons canviava la idea del data set que volíem anar generant, ha anat canviant el sistema d'etiquetatge de les diverses oracions que teníem.

Inicialment es va plantejar un sistema d'etiquetatge que anava del -1 al 1, on el -1 corresponia a oracions molt negatives i l'1 a oracions molt positives. El problema d'usar aquest sistema venia donat per la subjectivitat d'aquest tipus d'etiqueta, tot i que es van intentar establir uns criteris objectius, al haver-hi valors tant pròxims les diferències eren ínfimes, afegir que la *accuracy* del

---

<sup>16</sup> Sociedad Española del Procesamiento del Lenguaje Natural

<sup>17</sup> Elsafoury, Fatma (2020), "Cyberbullying datasets", Mendeley Data, V1, doi: 10.17632/jf4pzyvnpj.1

classificador tampoc era elevada doncs havia d'encertar correctament fins a 20 etiquetes diferents.

El següent plantejament va ser reduir aquestes etiquetes a 5 que corresponien de molt agressives a molt positives amb intervals de 0.5, es a dir: de -1 a -0.6 era molt agressiu, de -0.5 a -0.1 agressiu, 0 neutre i la resta a la inversa. Tot i això es va decidir que no eren necessàries les etiquetes que identificaven si una oració és positiva o no, i que ens desviàvem una mica de l'objectiu final el qual era determinar a nivell d'oració si s'està fent ciberbullying o no. Amb aquesta idea en ment, decidim finalment implementar un sistema d'etiquetatge que defineix la polaritat amb una variable dicotòmica, o 0 o 1, sent 0 una oració de ciberbullying i 1 una oració que no.

#### 10.1.4 Repercussions dels diferents intents

La cerca per a trobar un data set de base que ens servis per el que buscàvem com l'elaboració dels datasets finalment descartats així com finalment la traducció manual de 7000 oracions i etiquetatge d'aquests, ha comportat una gran major quantitat d'hores de les esperades i estimades quan es va dur a terme la organització de temps d'aquest projecte. Això ha comportat un endarreriment considerable, gairebé del doble d'hores de les estimades en la planificació temporal del treball que com s'explica i avalua en el punt 3.4 sobre riscos, ha comportat l'eliminació del tercer objectiu del projecte, aplicar-ho a una xarxa social ja existent, part que ha quedat pendent.

#### 10.2 Pre processat de les dades:

Una vegada generat el data set, processem les dades de manera que quedin el més homogènies possible i eliminem la màxima quantitat de dades innecessàries que només dificultarien l'aprenentatge del programa.

***Que follin a en Justin. És un imbècil de merda un ximple idiota! Kay? Bé!!!!!!!!!!!!***

*Figura 16: Exemple d'una oració inicial que tenim en el dataset*

##### 10.2.1 Tokenització

Consisteix en el procés mitjançant el qual transformem les nostres seqüències de paraules, es a dir, les nostres oracions en un llistat de paraules que seràn els tokens de manera que seràn aquests les entrades que rebrà l'algorisme.

Per a dur a terme la tokenització del text ens trobem amb diferents llibreries ja orientades a dur a terme aquesta tasca amb l'objectiu del processat natural del llenguatge. D'entre les llibreries més usades es troba la llibreria de nltk la qual et permet mitjançant la funció `word_tokenize(frase)` dur a terme aquesta acció fàcilment.

## 10.2.2 Eliminació de stopwords

Aquest procés consisteix en eliminar de la oració totes aquelles paraules que serveixen per a que la oració tingui sentit per a nosaltres sintàcticament, però que a l'hora de la veritat no varien el significat de la frase de manera que si les eliminem ens quedem amb les paraules determinants que ens permeten saber el sentiment d'una oració sense que aquest anàlisi sigui influenciat per a paraules sense rellevància. Un exemple d'aquest tipus de paraules serien els articles o els connectors.

Per a fer aquest procés es genera un llistat de paraules que son les anomenades *stopwords*, i es processen totes les paraules de les oracions, si es troben dins d'aquest llistat, son eliminades.

En altres llengües com l'anglès, els algorismes i llibreries enfocades al processat natural del llenguatge estan més avançades, i per tant, actualment existeix un llistat de paraules estandarditzat que et permet determinar quines de les paraules son les *stopwords*. Aquest no és el cas del català el qual després de cercar hem trobat només el llistat Generdiazjr<sup>18</sup> en Github per a usar, o un altre elaborat per el Luís de Yzagurre i Maura<sup>19</sup>. Després d'analitzar els dos llistats obtinguts i descobrir que contenien diferents *stopwords*, hem decidit elaborar el nostre llistat de paraules, de manera que sigui el més complet possible amb l'objectiu d'eliminar el màxim de soroll possible.

El llistat final es troba en l'**annex A** i està format per 779 paraules.

***follin Justin. imbècil merda ximple IDIOTA! Kay? Bé!!!!!!!!!!!!***

*Figura 17: Exemple oració del data set sense les stopwords*

## 10.2.3 Puntuació

Per a seguir polint el nostre text, eliminem també els signes de puntuació com punts, comes o signes d'interrogació. Per a fer-ho, usem altra vegada la llibreria nltk que te ja per si una part dedicada als signes de puntuació. Aquesta llibreria esta enfocada a l'anglès, per tant faltaria la detecció de, per exemple, el signe d'interrogació obert “¿”. Això però, no és un problema, doncs els signes que usem en català son els mateixos que usen en l'anglès, i per tant ja amb la llibreria els abarca tots.

Afegir que també aprofitem per afegir altres caràcters que volem que siguin eliminats, en concret com el data set esta format per oracions extretes de tweets, ens enfoquem en els signes més usats en la xarxa social que son “#” per als hashtags, i “@” per a les mencions.

***follin Justin imbècil merda ximple IDIOTA Kay Bé***

*Figura 18: Exemple oració del data set sense els signes de puntuació*

---

<sup>18</sup> <https://github.com/stopwords-iso/stopwords-ca/blob/222bb5691f90586016c1a3a8342e199ce7b3e399/raw/stop-words-catalian.txt>

<sup>19</sup> Luís de Yzagurre i Maura, director del Laboratori de Tecnologies Lingüístiques de l'IULA-UPC (Institut de Lingüística Aplicada de la Universitat Pompeu Fabra)

#### 10.2.4 Uppercase & lowercase

Seguint amb els processos d'homogeneïtzar les paraules, convertim totes les lletres de les paraules en lletres minúscules i eliminem els accents que s'hi puguin trobar de manera que una paraula que estigui escrita al inici d'una oració, i per tant amb majúscula, quedarà idèntica a la mateixa paraula en mig de la oració i no seran diferenciades.

***follin justin imbecil merda ximple idiota kay be***

*Figura 19: Exemple oració del data set tot en minúscules*

#### 10.2.5 Stemming & lematizació

Apliquem a les dades la operació de Stemming i lematització la qual consisteix en truncar la paraula a la seva arrel de manera que ens trobem amb el mateix "Lema" en totes les possibles transformacions d'una paraula. Un exemple seria la paraula: Cantaria, cantaré, cantant, i cantéssim, els quals serien tots reduïts a cantar, de manera que reduïm el nombre de paraules que l'algorisme es pot trobar, ja que el significat d'aquetes al ser totes variacions de la mateixa paraula, és el mateix.

Per a fer-ho, hem trobat dues possibles opcions que acceptin el llenguatge català. La primera és la eina de *Freeling*, la qual esta desenvolupada per un grup d'investigació de la UPC i que proporciona la opció de fer diferents accions semàntiques i morfològiques a una paraula i el que ens interessa a nosaltres, accions de lematització i stemmatització. La llibreria però, està escrita en C++ , i per a usar-la en python es necessita una API que permeti poder la usar amb el codi en python. A la vegada, hi ha una altra eina que consisteix en un algoritme de stemmatització generat per Israel Olalla com a part de l'empresa iSOCO i que permet el seu ús gratuït. Aquest codi ha estat inclòs en la llibreria Snowball la qual funciona perfectament amb python i ha estat provada la seva efectivitat segons la documentació, amb la pròpia eina de Freeling. Finalment ens decantem per aquesta segona opció doncs ens ofereix una major facilitat per a treballar amb ella.

***follar justar imbecil merda ximple idiota kay be***

*Figura 20: Exemple oració després stemming i lemmatization*

#### 10.2.6 Count vectorize

Finalment, una vegada hem obtingut el text en el format desitjat, l'estructura, de manera que quedi continguda en una matriu on hi ha tantes columnes com a paraules diferents en el data set, i per a cada oració, que correspon a les files de la matriu, s'estableix un numero el valor del qual correspon amb el nombre de vegades que la paraula surt a la oració. Aquesta matriu final és la que s'usa per a entrenar el nostre model. Per a aconseguir aquesta matriu la hem generat utilitzant la classe CountVectorizer de la llibreria Scikit-Learn, la qual et genera automàticament aquesta matriu amb les dades rebudes i els valors corresponents, és en la crida d'aquesta funció que la modifiquem per a que apliqui tot el preprocessat explicat en els punts anteriors, de manera que sigui eficient doncs les dades no es passen dues vegades sinó que per cada oració es preprocessa i es genera la seva fila a la matriu.

|                | follar | justar | imbecil | ... |
|----------------|--------|--------|---------|-----|
| Oració exemple | 1      | 1      | 1       | ... |
| Exemple 2      | 0      | 0      | 1       | ... |

Figura 21: Exemple taula final de la oració.

### 10.3 Generació del model

La elecció del classificador és fonamental en la cerca d'uns bons resultats de predicció. Sovint en els problemes de Machine learning i deep learning, es duen a terme proves amb diferents tipus de classificadors i hiperparàmetres fins a trobar aquells que siguin més òptims. De cara a trobar aquest classificador, ens plantejem dos possibles solucions, la primera és crear el nostre propi classificador, i la segona, és provar amb alguns dels classificadors ja proporcionats per la llibreria scikit-learn. Finalment s'escolliria el model que donés millors resultats.

#### 10.3.1 Creació del nostre propi classificador

Per a generar el que serà el nostre model decidim usar la llibreria Keras amb tensorflow per a intentar construir la nostra pròpia xarxa neuronal artificial que ens permeti classificar diferents texts com a ciberbullying o no.

El tipus de model que utilitzarem serà el seqüencial. Aquest és el més comú en aquest tipus de problemes i consisteix en un sistema on les dades es processen capa per capa fins a arribar al tipus de sortida desitjat, en el nostre cas, el vector és de dos sortides, 0 si és ciberbullying o 1 si no ho és, per tant usarem la funció d'activació de la sigmoide que ens donarà un dels dos resultats.

El model seqüencial estarà format per sis capes: la primera és la capa d'entrada o input layer, i consisteix en una capa amb les neurones que representen les dades inicials d'entrada. La segona capa, és la primera que definim amb Keras i consisteix en la capa d'Embedding, aquesta s'encarrega d'ajustar la capa d'entrada perquè sigui mapejada en el numero de neurones que se li indica i per tant rep: el numero de neurones, la funció d'activació i la dimensió de les dades d'entrada que en el nostre cas és 1.

Aquesta capa d'Embedding correspon a la primera capa ja oculta, la segona capa oculta i tercera en el model serà una capa Bidireccional que implementarà LSTM. Aquest tipus de capes són comunament usades en el processat del llenguatge natural ja que son xarxes recurrents que poden aprendre dependències a llargues distàncies. En aquest nivell haurem passat de 64 neurones a la capa d'embedding a 32 per la capa de LSTM i seguidament ho reduïm a 16 en la següent capa la qual serà una dense layer, o capa densa que consisteix en capes que connecten totes les seves sortides amb totes les de la capa anterior i per tant com cada neurona rep informació de totes les anteriors, es considera densament connectada. Entre aquesta capa i la següent, fem una capa de drop-out del 0.2 de manera que s'eliminen aleatòriament neurones de capes internes per a evitar situacions de overfitting. Finalment ens trobem amb la ultima capa del model, una dense layer la qual ens acaba donant un de dos valors aplicant la funció d'activació sigmoide, o 0 o 1.

```

Model: "sequential"
-----
Layer (type)                Output Shape              Param #
-----
text_vectorization (TextVect (None, None)              0
-----
embedding (Embedding)       (None, None, 64)         1920
-----
bidirectional (Bidirectional (None, 64)                 24832
-----
dense (Dense)                (None, 16)                1040
-----
dropout (Dropout)           (None, 16)                 0
-----
dense_1 (Dense)              (None, 2)                  34
-----
Total params: 27,826
Trainable params: 27,826
Non-trainable params: 0
-----

```

Figura 22: Resum del model seqüencial implementat

### 10.3.1.1 Entrenament del model

Una vegada hem creat aquest model, hem de decidir els paràmetres amb els que farem el training. En concret definim:

- **Funció d'optimització:** Consisteix en millorar els valors dels paràmetres per a reduir l'error de la xarxa neuronal. Després d'investigar, descobrim que el més recomanat per a aquest tipus d'aplicacions és el adam, i és amb el que ens decanem.
- **Funció de pèrdua (loss):** Consisteix en una funció que mesura l'error de predicció del model. En el nostre cas com consisteix en un mètode de classificació binaria, es recomana usar la funció de *binary crossentropy* que és per la que ens decanem.
- **Mètriques:** Consisteix en els paràmetres que volem controlar durant l'entrenament del model. Escollim el control de la *accuracy* que ens permet identificar la efectivitat del model fins a un cert punt.

En l'entrenament també determinem que l'ús de les dades serà com és habitual amb X identificant les dades d'entrenament, en el nostre cas les oracions i amb Y com les etiquetes del data set és a dir els 0 o 1. Les dades d'entrenament i les dades de

També després de provar diferents batch sizes, que consisteix en el nombre d'exemples que rep l'algorisme cada iteració en la fase training, decidim quedar-nos amb un batch\_size de 128. I sobre el nombre de epochs (les iteracions), en fem 30.

### 10.3.2 Classificadors de la llibreria scikit-learn

La llibreria scikit-learn ofereix uns 10 tipus diferents de classificadors, els quals poden ajustar els seus paràmetres per a una millor actuació en el model. En el nostre cas avaluem 5 d'aquests 10 classificadors que són els que millor s'adapten al nostre model i data set. Els cinc models són els explicats a continuació:

- **Linear SVM:**  
En aquest cas, el classificador genera una línia que s'encarrega de separar les dades en dos classes. Una vegada rebut el paràmetre a avaluar, si aquest cau a un costat o a l'altre de la línia rebrà l'etiqueta corresponent.
- **Kneighbors classifier:**  
El classificador realitza un mapeig de les dades d'aprenentatge. Una vegada obtinguda una data a analitzar, aquesta queda mapejada en el mateix pla i es miren el número K de veïns més propers que tingui el punt. El valor de K és imparell de manera que en cas de que no tots els valors siguin iguals s'etiquetaria la variable amb la etiqueta que tinguessin la majoria de K veïns més propers.
- **Decision tree classifier**  
L'objectiu d'aquest model és poder decidir un camí que porti a la decisió correcta basant-se en l'aprenentatge de regles de decisió que genera a partir de les dades de training, de manera que aprèn a agafar un camí un altre segons les dades obtingudes arribant finalment a una predicció.
- **Random Forest**  
És un algorisme de decisió basat en el mateix sistema que el decision tree classifier però amb un element aleatori de manera que el camí que es tria, ve donat o bé per les característiques que s'han obtingut o bé per un número aleatori, de manera que disminueix la tendència de overfitting que solen tenir els algorismes de decision tree.
- **MLP Classifier**  
Aquest classificador, a diferència dels altres i similar al nostre generat, conté internament una xarxa neuronal que és la que el porta a la decisió final de classificació, la gran diferència però, és que aquesta xarxa neuronal compta amb 100 capes ocultes, tot i que el funcionament és el mateix al ja explicat.

### 10.3.2.1 Entrenament dels models

Per a entrenar el model amb els diferents classificadors busquem que aquests estiguin programats amb els hiperparàmetres més adients per una actuació més òptima i un millor resultat. Per a obtenir aquests hiperparàmetres fem una grid search utilitzant la funció *GridSearchCV* de la llibreria scikit-learn la qual s'encarrega de fer l'entrenament provant tots els diferents valors possibles dins d'un interval per cada hiperparàmetre i finalment retornar el millor valor d'aquests de manera que ajustant els paràmetres als valors obtinguts, el classificador farà la millor actuació possible amb el data set rebut.

Executar aquesta funció és costosa de temps, en el nostre cas d'una durada superior a l'hora i mitja per cada classificador amb la que provis els paràmetres, però molt eficient doncs t'assegura que el rendiment que obtinguis amb el classificador és el màxim òptim que pots aconseguir amb el data set que tens per entrenar-lo.

## 11. Resultats: avaluació i optimitzacions

Una vegada hem generat el data set i tenim els classificadors, passem a la fase de training on entrenarem els nostres classificadors amb una part del data set que serà la part de train i un altre percentatge del data set servirà per a provar que el model funciona correctament. En concret dividim el data set de manera automàtica amb la funció *train\_test\_split* de la llibreria NOSEQUE i concretem que un 70% de les dades del dataset serviran per a entrenar el nostre model i un 30% per a provar-lo. A continuació expliquem el procés que s'ha dut a terme per a avaluar els nostres models així com a millores que hem implementat per a incrementar el rendiment d'aquests i les mètriques usades per avaluar-los.

### 11.1 Mètriques d'avaluació:

En concret per a determinar si un model és o no és suficientment bo mirem els següents paràmetres:

- **Accuracy** : Mesura el percentatge de vegades en les que el model ha encertat la resposta.
- **Sensitivity**: Mesura el percentatge de casos positius, és a dir 1, que han estat predits com a positius, per tant, positius veritaders
- **Specificity**: Mesura la proporció de casos negatius, és a dir 0, que han estat predits com a negatius, per tant, negatius veritaders.
- **False positive rate**: Mesura el percentatge d'error en les prediccions a 1, és a dir de totes les oracions predites a 1, quantes no ho eren, és a dir, falsos positius.
- **Precision** Indica el percentatge d'encert que tindrà el model quan prevegi que no és cyberbullying, és a dir, el percentatge de quantes vegades al etiquetar una oració com a 1, és 1.
- **ROC Curve**: Representa la proporció de positius veritaders vs. positius falsos en diferents llindars de classificació de manera que et permet calcular el valor ideal per a un major percentatge d'encerts.
- **Confusion matrix**: ens mostra de forma explícita les prediccions dutes a terme per el classificació de manera que puguem determinar els falsos positius i negatius i per tant, quan una classe es confon amb una altra.

### 11.2 Procés d'avaluació:

Inicialment hem testejat el model creat per a nosaltres doncs era el que més interès teníem en que funcionés. Uns resultats inicials son poc esperançadors, poc més d'un 30% d'accuracy, valor que aconseguim pujar fins el 35% després d'ajustar una mica els híperparàmetres dels models. Al ser aquest un resultat insuficient, ja que la tasa d'encert és inferior a llençar una moneda per endevinar el resultat (50%), provem de millorar el model. Després de parlar amb diverses persones més enteses en aquest àmbit, treiem les conclusions de que el model si més no és acceptable, i que el problema pot vindre donat per el data set.

Per a comprovar si els mals resultats són deguts a que el nostre data set és de baixa qualitat, es descarrega un data set amb varies opinions de ser usat anteriorment i que serveix per predir la polaritat (positiu o negatiu) d'un text en anglès. Una vegada provat amb aquest data set, el mateix model amb els mateixos hiperparàmetres ens dona una accuracy d'un 82%. Cosa que ens porta a intentar millorar el data set per a tindre un model funcional.

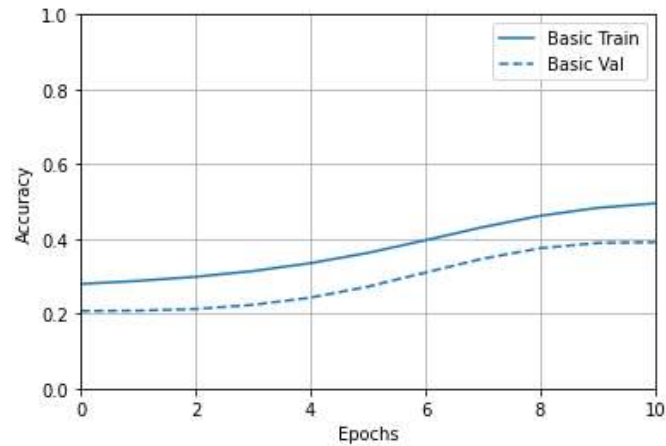


Figura 23: Gràfica de la accuracy en l'entrenament del model

Després de millorar el data set al màxim de les nostres possibilitats tenint en compte les constriccions de temps en les que ens trobem, aconseguim finalment una accuracy d'un 57% que tot i que es queda lluny d'un resultat elevat com el obtingut amb un altre data set, és una gran millora dels valors inicialment obtinguts .

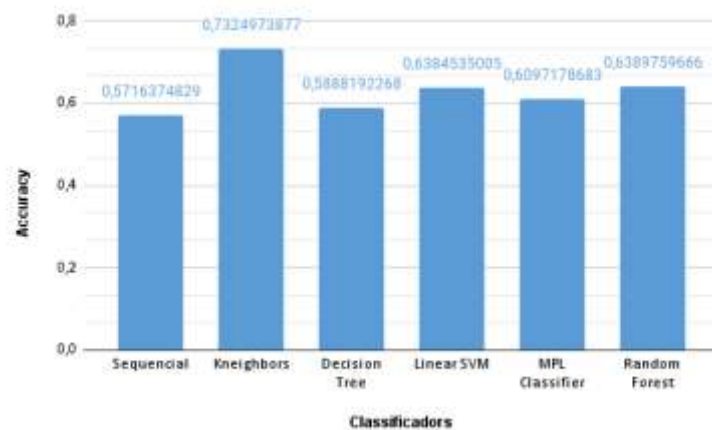


Figura 24: Gràfica comparativa de la accuracy dels diferents models

Passem doncs a provar amb els diferents classificadors ja existents, ara conscients de la limitació que la qualitat del nostre data set pot provocar. Després d'avaluar els cinc models examinats, cadascun amb els seus paràmetres més òptims calculats, es determina que el classificador que dona millors resultats és el KNeighbors Classifier amb una accuracy d'un 73%.

### 11.3 Anàlisi classificador escollit:

El model pel que s'ha optat finalment és el que usa el classificador de KNeighbors en concret amb una  $K = 3$ . A continuació mostrem les diferents mètriques d'aquest classificador que és el que ha donat uns millors valors de predicció amb les mètriques analitzades.

|                     |                     |
|---------------------|---------------------|
| Accuracy            | 0.7324973876698014  |
| Null Accuracy       | 0.5271682340647859  |
| Sensitivity         | 0.7569060773480663  |
| Specificity         | 0.7106045589692765  |
| False Positive Rate | 0.28939544103072345 |
| Precision           | 0.7011258955987717  |

Figura 25: Taula amb els valors de les mètriques analitzades

La null accuracy es correspon amb el valor d'etiquetes de testing que son 0 en el data set de prova, de manera que podem veure que el testing esta fet amb un conjunt relativament balancejat, amb un 52 % de frases de cyberbullying i un 48% de frases neutres. Més enllà, el percentatge de falsos negatius i falsos positius és relativament baix enfocant sobretot de cara a la gran quantitat d'encerts que es produeixen, dades també comprovables en la matriu de confusió del model.

Sobre la specificity i la Sensitivity, són valors relativament bons ja que els dos estan enfocats cap a l'1 de manera que un ajust dels hiperparametres i del threshold no es necessari ja que incrementaria un dels valors però disminuiria l'altre al ser inversament proporcionals, i actualment tenim els dos valors relativament elevats sobretot de cara a la corba ROC que s'ha obtingut.

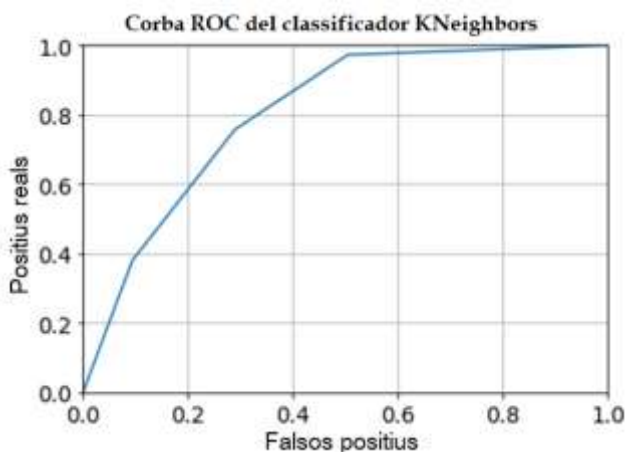


Figura 26: Corba ROC del model Kneighbors escollit

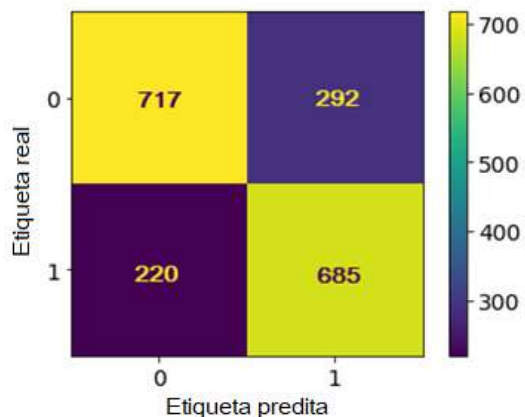


Figura 27: Matriu de confusió del model Kneighbors escollit

## 12. Conclusions

Finalment per a tancar el projecte s'exposa amb l'objectiu d'unificar els diferents temes exposats durant el treball un resum i conclusions finals a les que s'ha arribat amb la realització d'aquest projecte, així com els principals obstacles en el desenvolupament d'aquests, i com en un futur podria millorar-se.

El ciberbullying és una problemàtica molt present en la actualitat, està a l'ordre del dia principalment per l'augment considerable en les noves generacions les quals es veuen submergides en el món de l'internet i en especial, les xarxes socials cada vegada més prematurament, i degut a la anonimat i impunitat que tenen els agressors, així com la dificultat de les víctimes de poder combatre-la. Tot i que ja hi ha implementades algunes eines per a intentar fer front a aquest assetjament, aquestes no han resultat tan eficients com cabria esperar i per tant es buscava en el desenvolupament del projecte, intentar donar un enfocament nou a aquesta problemàtica.

El processat del llenguatge natural com a opció efectiva per a generar algoritmes és un entorn relativament recent a la llengua però amb una gran potència i espai per a créixer i desenvolupar. El projecte agafava un enfocament ambiciós i buscava la detecció de missatges de ciberbullying utilitzant aquest àmbit, en un llenguatge com és el català el qual no està avançat en aquesta disciplina.

Durant el desenvolupament del treball del qual es partia amb una desconexió gairebé total del entorn i matèria en el qual es volia desenvolupar l'eina, s'han adquirit extensos coneixements tant del processat del llenguatge natural -objectius i funcionament-, com de la creació de models de deep learning: com funcionen, com crear-ne, diversos usos i com objectivament avaluar-los. Aquests coneixements adquirits s'han posat en pràctica en la creació d'un model de deep learning, l'ús de models ja existents i la avaluació objectiva mitjançant mètriques determinades de cara a la creació d'un classificador que determinés si un missatge anava dirigit amb intencionalitat d'assetjament.

Al posar en pràctica aquests coneixements ha mostrat diferents problemes i matisos que no s'havien considerat com la dificultat i complexitat de la creació d'un data set adient per la tasca a satisfer o la sensibilitat dels models generats a una petita variació dels seus paràmetres o del entrenament que es du a terme.

Considerem doncs que l'objectiu del projecte, que era la creació d'un algorisme de deep learning que pogues detectar missatges de ciberbullying perquè es puguin realitzar les accions pertinents sobre aquests, s'ha complert. Tot i que inicialment a la planificació inicial es contemplava posar en pràctica aquest algorisme aplicant-lo a alguna situació real, finalment les constriccions temporals baix les que es desenvolupava el projecte ho han fet impossible.

El resultat si més no, és satisfactori ja que era un projecte ambiciós el qual era desenvolupat per una autora sense coneixença prèvia ni domini en el sector i no comptava amb material o suport generat prèviament en català sobre el que recolzar-s'hi. Tot i això s'ha aconseguit complir amb l'objectiu principal si més no hi queda espai per a millores i seguir avançant.

## 12.1 Problemes que han sorgit

Durant el desenvolupament del projecte han sorgit diversos problemes, alguns de menor rellevància i d'altres que han afectat considerablement el avanç del projecte. Principalment l'imprevist que ha resultat en el major endarreriment en l'avanç del treball ha sigut la creació del data set. La falta de corpus ja creats en català i la gran mida que es requereix per a poder entrenar uns classificadors d'aquest tipus no han sigut l'únic problema. En la creació del data set no contàvem amb la dificultat de trobar oracions enfocades en el sector que buscàvem, és a dir, oracions d'assetjament en català, i feia falta també comptar amb la necessitat de tindre una morfologia variada de manera que el model no s'esbiaixés i que aquest també estigues balancejat. Tot això sumat a la necessitat de traducció manual per a les incongruències sintàctiques que venen donades per el llenguatge col·loquial online, van provocar no únicament un endarreriment molt gran en la planificació temporal del projecte ja que van generar una situació bloquejant, sinó que també van tindre un efecte en la qualitat del data set amb el que s'ha treballat així com els resultats que s'han obtingut.

A aquest problema principal, se li sumen les dificultats per a trobar bones eines per a el pre-processat de les dades en català, en especial el procés de lematització i stemming, ja que per cerques online només es va trobar una eina no testada i per a saber de Freeling va requerir de la reunió amb un professor especialitzat en aquest camp que m'hi va informar, i els llargs temps d'execució i pes computacional que va tindre el training i cerques de hiperparàmetres pels diferents models, en el portàtil sobre el qual es treballava al qual li costava dur a terme aquestes tasques. En conclusió, diríem que la majoria de problemes són solucionables si treiem de la equació les constriccions de temps de les quals el projecte esta dotat.

## 12.2 Futures millores i continuació del projecte

Com hem comentat prèviament, degut a la inexistència d'un data set per a aquest tipus de treballs d'entrenament en català, el data set ha de ser generat per cadascú. De cara a continuar treballant amb el projecte i millorar-lo, es podria buscar d'ampliar el data set potser fins i tot buscant la recopilació de missatges reals de gent que ha viscut ciberbullying per a obtindre un data set més extens i realista.

També de cara a posar a punt l'algorisme generat com una eina funcional, es podria buscar salvar el percentatge d'encert obtingut el qual és d'un 73%, mirant no únicament un missatge sinó un conjunt de missatges d'una conversació. Ja que l'algorisme et retorna la categoria predita i també el percentatge de "seguretat" d'aquesta predicció, es podria buscar fer una suma dels percentatges d'un grup de missatges en una conversació i determinar segons el percentatge obtingut, si una situació de ciberbullying esta succeint en la conversació. D'aquesta manera es reduiria el percentatge d'error en la predicció, i s'aplicaria l'algorisme en un entorn més realista doncs les situacions d'assetjament no solen succeir únicament en un missatge.

També, anant més lluny i de cara a un objectiu final, es podria buscar aplicar aquest algorisme sobre converses de xarxes socials i fins i tot arribar a realitzar alguna acció automatitzada com seria el bloqueig d'un usuari de manera que l'eina arribaria ja a complir el seu objectiu final.

## Bibliografía

### Referències GEP:

- BullStop - Welcome to a safer Internet. Bullstop.io. Retrieved February 2021, from <https://www.bullstop.io/>.
- Datos sobre Bullying y Cyberbullying o acoso digital en España. Blog Educación y Bienestar digital. (2017). Retrieved 2 March 2021, from <https://gaptain.com/blog/bullying-ciberbullying-acoso-espana/>.
- De Castañeda, A. (2018). El cyberbullying en España. Zonamovilidad.es. Retrieved February 2021, from <https://www.zonamovilidad.es/ciberbullying-en-espana-iii-informe-fundacion-anar-fundacion-mutua-madrilena>.
- Ethnologue: Languages of the World. Ethnologue. Retrieved February 2021, from <https://www.ethnologue.com/>.
- Ganttter | #1 Cloud-Based Project Management Software. Ganttter. Retrieved February 2021, from <https://www.ganttter.com/>.
- jimcrickapp. <https://jimcrickapp.com/>. (2021). Retrieved February 2021, from <https://jimcrickapp.com/>.
- Los idiomas, en cifras: ¿cuántas lenguas hay en el mundo?. europapress.es. Retrieved February 2021, from <https://www.europapress.es/sociedad/noticia-idiomas-cifras-cuantas-lenguas-hay-mundo-20190221115202.html>.
- Moreno, A. Procesamiento del lenguaje natural ¿qué es? - IIC. Instituto de Ingeniería del Conocimiento. Retrieved February 2021, from <https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/>.
- Noriega Cabrera, L. Ganttter | #1 Cloud-Based Project Management Software. Ganttter. Retrieved February 2021, from <https://www.ganttter.com/>.
- ¿Qué es Aprendizaje profundo (deep learning)? - Definición en WhatIs.com. SearchDataCenter&nbsp;en&nbsp;Español. (2017). Retrieved February 2021, from <https://searchdatacenter.techtarget.com/es/definicion/Aprendizaje-profundo-deep-learning>.
- Rojas, E. (2018). Glosario de los seis términos básicos del Machine Learning. MuyComputerPRO. Retrieved February 2021, from <https://www.muycomputerpro.com/2018/02/07/glosario-terminos-basicos-machine-learning>.

- Tecnologies de la informació | Facultat d'Informàtica de Barcelona. Fib.upc.edu. (2021). Retrieved February 2021, from <https://www.fib.upc.edu/ca/estudis/graus/grau-en-enginyeria-informatica/pla-destudis/especialitats/tecnologies-de-la-informacio>.
- Trello. Trello.com. Retrieved February 2021, from <https://trello.com/>.
- WordPress.com: crea un sitio web o blog gratuito. WordPress.com. Retrieved February 2021, from <https://wordpress.com/es/>.

### Sobre el Dataset:

- Balaganur, S. (2021). 10 Popular Datasets For Sentiment Analysis. Retrieved 28 January 2021, from <https://analyticsindiamag.com/10-popular-datasets-for-sentiment-analysis/>
- Best Spanish Language Datasets for Machine Learning | Lionbridge AI. (2021). Retrieved 17 January 2021, from <https://lionbridge.ai/datasets/22-best-spanish-language-datasets-for-machine-learning/>
- Catalan. (2021). Retrieved 1 May 2021, from <https://www.ranks.nl/stopwords/catalan>
- Catalan Stop Words. (2021). Retrieved 17 March 2021, from [http://latel.upf.edu/morgana/altres/pub/ca\\_stop.htm](http://latel.upf.edu/morgana/altres/pub/ca_stop.htm)
- ChatCoder Data page. (2021). Retrieved February 2021, from <https://www.chatcoder.com/data.html>
- Datasets - Linked Data Models for Emotion and Sentiment Analysis Community Group. (2021). Retrieved 3 February 2021, from <https://www.w3.org/community/sentiment/wiki/Datasets>
- Elsafoury, F. (2021). Cyberbullying datasets. Retrieved 9 March 2021, from [https://data.mendeley.com/datasets/jf4pzyvnpj/1#\\_\\_sid=js0](https://data.mendeley.com/datasets/jf4pzyvnpj/1#__sid=js0)
- nltk.tokenize.punkt — NLTK 3.6 documentation. (2021). Retrieved 8 June 2021, from [https://www.nltk.org/\\_modules/nltk/tokenize/punkt.html](https://www.nltk.org/_modules/nltk/tokenize/punkt.html)
- Sentiment Analysis to Detect Threatening Tweets in a Collaborative Team. (2021). Retrieved 30 April 2021, from <https://medium.com/omdena/sentiment-analysis-to-detect-threatening-tweets-in-a-collaborative-team-b6ee11ffe179>
- Sentiment — polyglot 16.07.04 documentation. (2021). Retrieved May 2021, from <https://polyglot.readthedocs.io/en/latest/Sentiment.html>
- snowballstemmer. (2021). Retrieved from <https://pypi.org/project/snowballstemmer/>

- stemmer. (2021). how to use snowball 's catalan stemmer?. Retrieved from <https://stackoverflow.com/questions/35946932/how-to-use-snowballs-catalan-stemmer>
- stop-words. (2021). Retrieved 15 January 2021, from <https://pypi.org/project/stop-words/>
- stopwords-iso/stopwords-ca. (2021). Retrieved 5 May 2021, from <https://github.com/stopwords-iso/stopwords-ca/blob/master/stopwords-ca.txt>
- Where can I access datasets on cyberbullying. (2021). Retrieved 11 April 2021, from [https://www.researchgate.net/post/Where\\_can\\_I\\_access\\_datasets\\_on\\_cyberbullying](https://www.researchgate.net/post/Where_can_I_access_datasets_on_cyberbullying)

### Sobre el Model:

- BERT-Sentiment-Analysis-Twitter-Spanish. (2021). Retrieved 17 January 2021, from <https://github.com/frantrucco/BERT-Sentiment-Analysis-Twitter-Spanish>
- ¿Cómo funciona la capa de 'incrustación' de Keras?. (2021). Retrieved 20 February 2021, from <https://qastack.mx/stats/270546/how-does-keras-embedding-layer-work>
- Cómo usar redes neuronales (LSTM) en la predicción de averías en las máquinas. (2021). Retrieved 2 June 2021, from <https://blog.gft.com/es/2018/11/06/como-usar-redes-neuronales-lstm-en-la-prediccion-de-averias-en-las-maquinas/>
- Conceptos Fundamentales en Machine Learning - Función de Perdida y Optimización. (2021). Retrieved 14 May 2021, from <https://planetachatbot.com/conceptos-fundamentales-en-machine-learning-funcon-de-perdida-y-optimizacion/>
- ¡Domina machine learning y computer vision en español!. (2021). Retrieved 11 March 2021, from <https://datasmarts.net/es/que-es-un-optimizador-y-para-que-se-usa-en-deep-learning/>
- Ebrahim, M. (2021). Tutorial de NLP con Python NLTK (ejemplos simples) - Like Geeks. Retrieved 5 April 2021, from <https://likegeeks.com/es/tutorial-de-nlp-con-python-nltk/>
- El modelo Embeddings (Incrustaciones) de Palabras - ▷ Cursos de Programación de 0 a Experto © Garantizados. (2021). Retrieved 28 May 2021, from <https://unipython.com/el-modelo-embeddings-incrustaciones-de-palabras/#:~:text=La%20capa%20de%20embeddings%20se,a%20los%20vectores%20de%20palabras>
- Garrido, M. (2021). Cómo hacer Análisis de Sentimiento en español – Pybonacci. Retrieved 23 March 2021, from <https://pybonacci.org/2015/11/24/como-hacer-analisis-de-sentimiento-en-espanol-2/>
- Heras, J. (2021). Precision, Recall, F1, Accuracy en clasificación - IArtificial.net. Retrieved 23 February 2021, from <https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/>

- ML Metrics: Sensitivity vs. Specificity - DZone AI. (2021). Retrieved 20 April 2021, from <https://dzone.com/articles/ml-metrics-sensitivity-vs-specificity-difference>
- NNmulticapa.md. (2021). Retrieved 19 April 2021, from [http://personal.cimat.mx:8181/~mrivera/cursos/aprendizaje\\_profundo/nn\\_multicapa/nn\\_multicapa.html](http://personal.cimat.mx:8181/~mrivera/cursos/aprendizaje_profundo/nn_multicapa/nn_multicapa.html)
- Realiza un análisis de sentimiento en 3 pasos con python. (2021). Retrieved 2 April 2021, from <https://platzi.com/tutoriales/1874-python-lenguaje-natural/5654-realiza-un-analisis-de-sentimiento-en-3-pasos-con-python/>

### Sobre les Llibraries:

- Alberca, A. (2021). La librería Pandas | Aprende con Alf. Retrieved from <https://aprendeconalf.es/docencia/python/manual/pandas/>
- Alberca, A. (2021). La librería Matplotlib | Aprende con Alf. Retrieved from <https://aprendeconalf.es/docencia/python/manual/matplotlib/>
- Decision Trees — scikit-learn 0.24.2 documentation. (2021). Retrieved from <https://scikit-learn.org/stable/modules/tree.html#tree>
- 1.11. Ensemble methods — scikit-learn 0.24.2 documentation. (2021). Retrieved from <https://scikit-learn.org/stable/modules/ensemble.html#forest>
- Examples — scikit-learn 0.24.2 documentation. (2021). Retrieved from [https://scikit-learn.org/stable/auto\\_examples/index.html#classification](https://scikit-learn.org/stable/auto_examples/index.html#classification)
- Group, E. (2021). Redes neuronales con Python: ¿por qué es el mejor lenguaje para IA?. Retrieved from <https://blog.enzymeadvisinggroup.com/redes-neuronales-python>
- Instalando StopWords en NLTK. (2021). Retrieved from <https://jantoniomora.wordpress.com/2017/08/22/instalando-stopwords-en-nltk/>
- Libreria Scikit Learn de Python. (2021). Retrieved from <https://aprendeia.com/libreria-scikit-learn-de-python/>
- Natural Language Toolkit — NLTK 3.6.2 documentation. (2021). Retrieved from <http://www.nltk.org/>
- Nearest Neighbors — scikit-learn 0.24.2 documentation. (2021). Retrieved from <https://scikit-learn.org/stable/modules/neighbors.html#classification>
- Nltk - Data. (2021). Retrieved from [http://www.nltk.org/nltk\\_data/](http://www.nltk.org/nltk_data/)
- (NLTK), N. (2021). The Natural Language Toolkit (NLTK) Open Source Project on Open Hub. Retrieved from <https://www.openhub.net/p/nltk>

- Programació en Python - 3. Natural Language Toolkit (NLTK). (2021). Retrieved from <https://sites.google.com/view/programacio-en-python/home/3-natural-language-toolkit-nltk>
- sklearn.neural\_network.MLPClassifier — scikit-learn 0.24.2 documentation. (2021). Retrieved from [https://scikit-learn.org/stable/modules/generated/sklearn.neural\\_network.MLPClassifier.html](https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html)
- sklearn.svm.LinearSVC — scikit-learn 0.24.2 documentation. (2021). Retrieved from <https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC>
- Snowball. (2021). Retrieved from <https://snowballstem.org/>
- Supervised learning — scikit-learn 0.24.2 documentation. (2021). Retrieved from [https://scikit-learn.org/stable/supervised\\_learning.html#supervised-learning](https://scikit-learn.org/stable/supervised_learning.html#supervised-learning)
- Support Vector Machines — scikit-learn 0.24.2 documentation. (2021). Retrieved from <https://scikit-learn.org/stable/modules/svm.html#svm-classification>
- Support Vector Machines(SVM) — An Overview. (2021). Retrieved from <https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989>

## Annexes

### Annex A: Stopwords

Llistat de stopwords generat, amb un total de 779 paraules en català com a resultat de la combinació de les diverses llistes existents.

- a
- abans
- abans-d'ahir
- abintestat
- ací
- adesiara
- adés
- adéu
- adàgio
- ah
- ahir
- ai
- aitambé
- aitampoc
- aitan
- aitant
- aitantost
- aixà
- així
- això
- al
- aleshores
- algú
- algun
- alguna
- algunes
- alguns
- alhora
- allà
- allèn
- allí
- allò
- almenys
- als
- alto
- altra
- altre
- altres
- altresí
- altri
- alça
- al·legro
- amargament
- amb
- ambdues
- ambdós
- amunt
- amén
- anar
- ans
- andante
- andantino
- anit
- antany
- apa
- après
- aqueix
- aqueixa
- aqueixes
- aqueixos
- aquell
- aquella
- aquelles
- aquells
- aquest
- aquesta
- aquestes
- aquests
- aquèu
- aquí
- ara
- arran
- arrere
- arreu
- arri
- arruix
- aquí
- atxim
- au
- avall
- avant
- aviat
- avui
- açò
- bah
- baix
- baldament
- ballmanetes
- banzim-banzam
- bastant
- bastants
- ben
- bis
- bitllo-bitllo
- bo
- bé
- ca
- cada
- cal
- cap
- car
- caram
- catorze
- cent
- centes
- cents
- cerca
- cert
- certa
- certes
- certs
- cinc
- cinquanta
- cinquena

- cinquenes
- cinquens
- cinquè
- com
- comsevulla
- contra
- cordons
- corrents
- cric-crac
- cadascuna
- cadascunes
- cadascuns
- cadascú
- com
- consegueixo
- consequim
- aconseguir
- consigueix
- consigueixen
- consigueixes
- contra
- d
- d'un
- d'una
- d'unes
- d'uns
- daixoneses
- daixò
- dalloneses
- dallò
- daltabaix
- damunt
- darrera
- darrere
- davall
- davant
- dalt
- de
- del
- dels
- des
- des de
- després
- dins
- dintre
- donat
- doncs
- debades
- dedins
- defora
- dejorn
- dejús
- dellà
- dementre
- dempeus
- demés
- demà
- des
- desena
- desenes
- desens
- després
- dessorbre
- dessota
- dessús
- desè
- deu
- devers
- devora
- deçà
- diferents
- dinou
- dins
- dintre
- disset
- divers
- diversa
- diverses
- diversos
- divuit
- doncs
- dos
- dotze
- dues
- durant
- ecs
- e
- eh
- el
- ela
- elis
- ell
- ella
- elles
- ells
- els
- em
- emperò
- en
- enans
- enant
- encara
- encontinent
- endalt
- endarrera
- endarrere
- endavant
- endebades
- endemig
- endemés
- endemà
- endins
- endintre
- enfora
- engir
- enguany
- enguanyasses
- enjús
- enlaire
- enlloc
- enllà
- enrera
- enrere
- ens
- ens
- ensems
- ensota
- ensús
- entorn
- entre
- entremig
- entretant
- entrò
- envers
- envides
- environs
- enviro
- ençà

- ep
- ep
- era
- erem
- eren
- eres
- es
- escar
- essent
- esser
- est
- esta
- estada
- estades
- estan
- estant
- estar
- estaran
- estarem
- estareu
- estaria
- estarien
- estaries
- estaré
- estarà
- estaràs
- estaríem
- estaríeu
- estat
- estats
- estava
- estaven
- estem
- estes
- esteu
- estic
- estiguem
- estigueren
- estigueres
- estigues
- estiguessis
- estigueu
- estigui
- estiguin
- estiguis
- estigué
- estiguérem
- estiguéreu
- estigués
- estiguí
- estos
- està
- estàs
- estàvem
- estàveu
- et
- etc
- etcètera
- ets
- excepte
- fa
- faig
- fan
- fas
- fem
- fer
- feu
- fi
- fins
- fora
- foren
- fores
- força
- fos
- fossin
- fosis
- fou
- fra
- fui
- fóra
- fórem
- fóreu
- fóreu
- fóssim
- fóssiu
- gaire
- gairebé
- gaires
- gens
- girientorn
- gratis
- ha
- hagi
- hagin
- hakis
- haguda
- hagudes
- hagneren
- hagneres
- haguessin
- haguessis
- hagut
- haguts
- hagué
- haguérem
- haguéreu
- hagués
- haguéssim
- haguéssiu
- haguí
- hala
- han
- has
- hauran
- haurem
- haureu
- hauria
- haurien
- hauries
- hauré
- haurà
- hauràs
- hauríem
- hauríeu
- havem
- havent
- haver
- haveu
- havia
- havien
- havies
- haviem
- haviéu
- he
- hem
- heu
- hi
- ho

- hom
- hui
- hàgim
- hàgiu
- i
- igual
- iguals
- inclòs
- inclusiu
- ja
- jamai
- jo
- l
- l'hi
- la
- leri-leri
- les
- li
- li'n
- lla
- llarg
- llavors
- llevat
- lluny
- llur
- llurs
- lo
- los
- ls
- m
- m'he
- ma
- mai
- mal
- malament
- malgrat
- manco
- mant
- manta
- mantes
- mantinent
- mants
- massa
- mateix
- mateixa
- mateixes
- mateixos
- me
- mentre
- mentrestant
- menys
- mes
- meu
- meua
- meues
- meus
- meva
- meves
- mode
- mi
- mig
- mil
- mitges
- mitja
- mitjançant
- mitjos
- moixoní
- molt
- molta
- moltes
- molts
- mon
- mos
- mons
- més
- n
- na
- n'he
- n'hi
- ne
- ni
- ningú
- no
- nogensmenys
- només
- noranta
- nos
- nosaltres
- nostra
- nostre
- nostres
- nou
- novena
- novenes
- novens
- novè
- nòs
- nós
- o
- oh
- oi
- oida
- on
- onsevulga
- onsevulla
- onze
- pas
- pengim-penjam
- pel
- pels
- per
- per que
- perquè
- pertot
- però
- piano
- pla
- poc
- poc
- poca
- pocs
- podem
- poden
- poder
- podeu
- poques
- potser
- prest
- primer
- primera
- primeres
- primers
- pro
- prompte
- prop
- prou
- puix
- pus

- propi
- pàssim
- puc
- qual
- quals
- qualsevol
- qualsevulla
- qualssevol
- qualssevulla
- quan
- quant
- quanta
- quantes
- quants
- quaranta
- quart
- quarta
- quartes
- quarts
- quasi
- quatre
- que
- quelcom
- qui
- quin
- quina
- quines
- quins
- quinze
- quisvulla
- què
- ran
- re
- rebé
- renoi
- rera
- rere
- res
- retruc
- s
- s'ha
- s'han
- sa
- sabem
- saben
- saber
- sabeu
- sap
- saps
- semblant
- semblants
- sense
- salvament
- salvant
- salvat
- se
- segon
- segona
- segones
- segons
- seguida
- seixanta
- sempre
- sengles
- sens
- sense
- ser
- seran
- serem
- sereu
- seria
- serien
- series
- seré
- serà
- seràs
- seríem
- seríeu
- ses
- set
- setanta
- setena
- setenes
- setens
- setze
- setè
- seu
- seua
- seues
- seus
- seva
- seves
- si
- sia
- siau
- sic
- siguem
- sigues
- sigueu
- sigui
- siguin
- siguis
- sinó
- sis
- sisena
- sisenes
- sisens
- sisè
- sobre
- sobretot
- soc
- sol
- sola
- solament
- soles
- sols
- som
- son
- sons
- sos
- sota
- sots
- sou
- sovint
- suara
- sí
- sóc
- són
- t
- t'ha
- t'han
- t'he
- ta
- tal
- tals
- també
- tampoc
- tan

- tanmateix
- tant
- tanta
- tantes
- tantost
- tants
- te
- tercer
- tercera
- terceres
- tercers
- tenir
- teniu
- teu
- teua
- teues
- teus
- teva
- teves
- tinc
- ton
- tos
- tons
- tost
- tostemps
- tot
- tota
- total
- totes
- tothom
- tothora
- tots
- trenta
- tres
- tret
- tretze
- tu
- tururut
- u
- uf
- ui
- uix
- ultra
- un
- una
- unes
- uns
- us
- up
- upa
- va
- vagi
- vagin
- vagis
- vaig
- vair
- vam
- van
- vares
- vas
- vau
- vem
- verbigràcia
- vers
- vet
- veu
- vint
- vora
- vos
- vosaltres
- vostra
- vostre
- vostres
- vostè
- vostès
- vuit
- vuitanta
- vuitena
- vuitenes
- vuitens
- vuitè
- vés
- vèrem
- vàrem
- vàreu
- vós
- xano-xano
- xau-xau
- xec
- érem
- éreu
- és
- ésser
- éssent
- àdhuc
- àlies
- ça
- ço
- òlim
- ídem
- últim
- última
- últimes
- últims
- únic
- única
- únics
- úniques.
- ús