Detecció de depressió a xarxes socials mitjançant Machine Learning

Martí Caixal i Joaniquet

Resum– Les xarxes socials actualment estan presents a la vida de moltes persones i, si bé aporten coses bones, també són un lloc que s'utilitza per parlar de la vida d'un mateix i de les seves desgràcies. Aquest treball busca avaluar i posar a prova els varis mètodes que hi ha per analitzar aquest tipus de comportament de les persones. L'objectiu principal és veure quins mètodes tenen més èxit per detectar els missatges i publicacions que indiquen que una persona pateix depressió. Aquests mètodes són molt importants per fer deteccions preventives de depressió i ansietat i per això cal veure el comportament de cada mètode, quines mancances té i quines avantatges.

Paraules clau– Intel·ligència Artificial, processat de llenguatge natural, Python, Aprenentatge computacional, aprenentatge profund, anàlisi de sentiments, depressió, preprocessament de dades

Abstract– Nowadays social networks are present in a lot of people's lives and, besides bringing joy to some, they are also a place where people express their misfortunes and tragedies. This project aims to evaluate and put into test the multiple methods that exist to analyse this kind of people behaviour. The main objective is to see which of this methods are more successful to detect messages and posts that would lead a professional to think that user suffers from depression. This methods are very important in order to be able to make preventive depression and anxiety detections, and thus it is needed to see what each method behaves like, its disadvantadges and advanages.

Keywords– Artificial Intelligence, Natural Language Processing, Python, Machine Learning, Deep Learning, Sentiment Analysis, Depression, Data Preprocessing

1 Introducció - Context del treball

Les xarxes socials són considerades com uns sistemes d'informació en línia que permeten compartir l'estil de vida dels seus usuaris. Cada un té el seu perfil personal on penja actualitzacions del seu dia a dia i la resta d'usuaris poden reaccionar-hi i posar comentaris. De la mateixa manera, també es permet seguir a gent i altres institucions que siguin de l'agrat d'un [1].

L'inici de les xarxes socials es remunta a finals dels anys 90, quan l'Internet tot just passava de ser una eina orientada a professionals a ser d'àmbit general. Tot i no ser la primera en aparèixer, la xarxa "MySpace" va ser la que va popularitzar aquest fenomen i va obrir pas a un seguit de noves xarxes socials. La més famosa, i que actualment segueix sent la que té més usuaris actius, és "Facebook", creada per en Mark Zuckerberg. Si bé inicialment l'únic objectiu era estar en contacte amb la gent del teu cercle més proper, avui en dia les xarxes socials són un mitjà per la gent famosa on rebre milers i milers de seguidors i visualitzacions, deixant de banda l'objectiu principal amb el que es van crear. [2]

Les xarxes socials no estan exemptes de problemàtiques. Al cap i a la fi, són un lloc on tothom pot dir la seva sense cap tipus de restricció. Això ha portat fins al punt on la gent diu allà el que no és capaç o bé no s'atreveix a dir en persona.

Sí bé aquestes xarxes tenen codis de conducta i disposen d'equips de moderadors, no es pot fer front a tots els problemes. Al fet que se li dona més importància i es destinen més recursos de forma activa és a l'anomenat "cyber-bullying", doncs és el que més destaca i no deixa de ser un atac des d'un individu cap a un altre. Un altre problema important per les conseqüències i el volum d'afectats, és la depressió. L'ús de les xarxes socials

• E-mail de contacte: mcaixal1999@gmail.com
• Menció realitzada: Computació
• Treball tutoritzat per: Ramon Baldrich (Ciències de la Computació)
• Curs 2022/23

possibilita que les persones amb inestabilitat emocional o amb depressió es puguin expressar i deixin anar les seves preocupacions. [3]

Un estudi demostra que més d'un 20% dels usuaris han penjat comentaris amb indicis que podrien estar patint depressió o similar. No només això, sinó que és una tendència que està en augment, havent-hi el doble de casos, proporcionalment, ara que fa 10 anys. Addicionalment, hi ha hagut alguns casos on els usuaris expliquen les "penúries" del seu dia a dia fins al punt on escriuen allà mateix la nota de suïcidi. [4]

Clarament, totes aquestes notícies han provocat un seguit de queixes a les empreses propietàries de les xarxes per part de moltes organitzacions i institucions. Els responsables de moderació de les xarxes socials es defensen dient que no hi ha manera de poder veure tots els posts amb indicis de depressió. A diferència dels que contenen "cyber-bullying" o similars, que són reportats per altres usuaris (normalment les víctimes), els missatges amb continguts depriments passen desapercebuts, o simplement no se'ls hi dona importància, per la resta d'usuaris.

Aquest fet dificulta moltíssim la feina dels equips moderadors, els quals no tenen els mitjans necessaris per avaluar tots els missatges i comentaris. La manca de normativa dificulta les actuacions i les possibles responsabilitats.

2 OBJECTIUS I ESTAT DE L'ART

Així doncs, hi ha la necessitat d'obtenir un sistema que pugui fer front al problema. Per la seva pròpia naturalesa, s'ha de solucionar no pas actuant un cop passa, sinó de forma preventiva prenent accions abans de que sigui massa tard.

L'objectiu és identificar els casos d'usuaris que necessitin ajuda mitjançant models predictius basats en intel·ligència artificial. Més específicament, cal treballar i investigar l'anomenat "Natural Language Processing" (NLP), traduït a processament de llenguatge natural.

Clarament, ja hi ha molts mètodes i models disponibles que realitzen la tasca desitjada. L'objectiu, per tant, no és crear des de zero un nou mètode, sinó fer un estudi de l'eficiència i l'èxit que tenen cadascun d'ells. Per tant, s'implementaran un seguit de mètodes diferents i es procedirà a fer les proves adients. Les dades utilitzades són "datasets"ja classificats correctament. Aquestes dades són extretes directament i sense tractar de les xarxes socials "Twitterï "Reddit". Aquest fet per una banda permet tenir una representació pràcticament exacta de les dades amb les que s'enfronten els varis models en el moment de la veritat. Per una altra banda, al ser informació sense tractar, també obre la porta a fer un "Exploratory Data Analysis"(EDA) i treure ja unes estadístiques i característiques preliminars, les quals després es podran comparar amb els resultats arribats un cop executats els models. Addicionalment, els models es posaran a prova tant amb les dades sense tractar, com fent un previ tractament del dataset amb l'objectiu de veure el nou comportament dels models i si hi ha algun indici de millora a les produccions.

Dins del "Machine Learning"(ML) ja s'ha fet un seguit d'investigacions i treballs des dels quals es basaran els objectius.

2.1 Mètodes tradicionals

Per una banda hi ha els mètodes tradicionals de ML. L'article "Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier" [5] explora tant "Naïve Bayes"(NB) i "K-nearest neighbours"(K-NN), intentant avaluar el rendiment i resultats de cada un. Tant un com l'altre es comporten de forma similar i es postulen com a bones opcions dins d'aquest tipus de mètodes, però sent el NB el que ho fa millor amb el 90% d'"accuracy". Tot i això, l'estudi d'aquests mètodes s'orienta a fer un "sentiment analysis"(SA) d'opinions de pel·lícules, no pas de posts a les xarxes socials. També hi ha més estudis, com el "Sentiment Analysis on Twitter Data using KNN and SVM"[6] que sí que fan el SA sobre un cas de xarxes socials. Per aquest últim, el mètode de "Support Vector Machine"(SVM) és el que dona més bons resultats. Tot i això, no s'està buscant específicament depressió, sinó si el "tweet"és positiu o negatiu.

Un dels objectius als que es vol arribar és comparar varis models tradicionals i posar-los a prova directament en detectar depressió als posts en xarxes socials. D'aquesta manera es podrà veure si els que tenen més èxit en SA genèric, també el tenen quan es busca un sentiment en concret. Més específicament, per aquesta part es posaran a prova els classificadors NB, "Decision Tree"(DT), "Random Forest"(RF), SVM i KNN.

Finalment, aquests mètodes més tradicionals tenen un seguit de paràmetres, anomenats "Hyper Parameters", amb els quals configurar el seu comportament. Segons l'article "Hyperparameter Search in Machine Learning" [12], aconseguir trobar els valors adequats pot marcar la diferència entre unes bones prediccions i unes males. Caldrà veure fins a quin punt una bona elecció de "Hyper Parameters"pot afectar a les prediccions quan es busca un SA de depressió.

2.2 Deep Learning

Per una altra banda, també hi ha els mètodes basats en el "Deep Learning"(DL) que actualment estan triomfant més. L'estudi "Comparing Machine Learning and Deep Learning Approaches on NLP Tasks for the Italian Language" [13] parla sobre les diferències que hi ha entre el DL i els mètodes tradicionals de ML, concloent que per NLP el ML no acaba de tenir superioritat en les tasques de classificació que depenent molt en l'anàlisi semàntic.

Un mètode que ja porta un temps utilitzant-se per NLP són les anomenades "Recurrent Neural networks"(RNN). Es diferencien de les xarxes neuronals convencionals, i són especialment útils en NLP, pel fet que el ïnput"és una sola paraula, donant flexibilitat per treballar amb diferents llargades a les frases. Hi ha dos models que fins fa poc eren

els més utilitzats, anomenats L·LSTMï ”GRU”. L'article “LSTM and GRU neural network performance comparison study” [17] explora aquests dos mètodes, on conclou que el fet de retenir informació els hi permet posicionar-se al podi de les RNN. No obstant, està comparant amb tasques de molts camps, sense entrar en detall al NLP i encara menys al SA. Un dels objectius serà aplicar ambdós mètodes per detectar depressió a les xarxes socials.

L'article “Attention is All you Need” [14], publicat per Google el 2017, presenta un nou model de xarxes neuronals especialment innovador a l'apartat de NLP. S'anomena “Transformer” i es basa processar dades d'entrada seqüencials però, a diferència de les Xarxes Neuronals Recurrents, processen tota l'entrada alhora. L'article “Overview of the Transformer-based Models for NLP Tasks”[16] fa especialment atenció en el rendiment dels models ”Transformer”amb els problemes NLP, concloent que efectivament han millorat molt les prediccions respecte a la resta de mètodes utilitzats fins al moment. Un objectiu serà aplicar models ”Transformer”(BERT) i veure com es comporten a l'hora de detectar casos de depressió a les xarxes socials, comparant-los amb la resta de mètodes provats anteriorment.

La figura 1 mostra un diagrama els mètodes que s'utilitzen al treball.

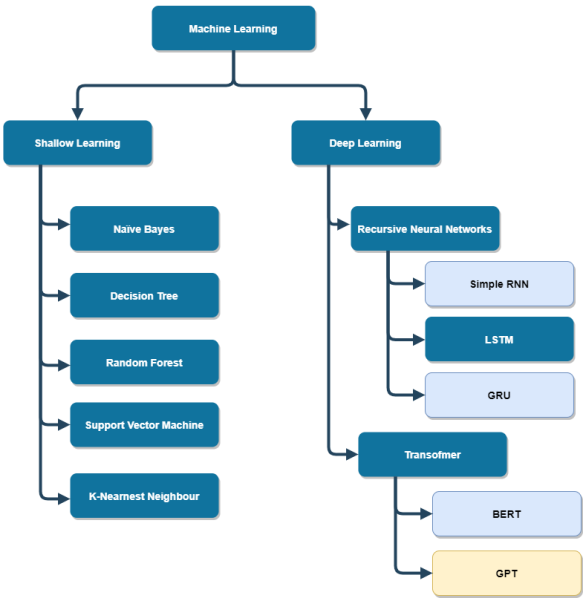

Fig. 1: Mètodes utilitzats en el projecte. En blau fort els que s'hi ha aprofundit. En blau fluix els que s'han fet per sobre. En groc els que no s'ha pogut fer.

3 Metodologia i desenvolupament

Aquesta secció explica com s'ha portat a terme el treball. Des de la planificació i les eines utilitzades, fins a les pautes per desenvolupar i guardar resultats.

3.1 Planificació

Per poder desenvolupar el projecte de manera adequada i ordenada, és necessari establir una metodologia de treball a seguir, que permeti tenir un bon control del flux de treball i compliment de les tasques i entregues. La metodologia escollida és l'anomenada “àgil”.

Es basa en iteracions curtes, adaptabilitat als canvis, facilitat de detectar i solucionar errors, i entregues parcials. És un projecte que es pot subdividir en diferents apartats independents entre ells i aquesta metodologia és especialment útil en aquests casos. D'aquesta manera, si una part presenta complicacions, la resta no es veuran pràcticament afectades. També permet fer, al final de cada iteració, un anàlisi de la situació actual, valorant el que s'ha realitzat, el que manca per realitzar i qualsevol aspecte que calgui modificar o adaptar.

Finalment, també es fa us d'un controlador de versions GIT, tenint així un registre i historial de tots els canvis i permetent revertir-los en cas de fer falta.

3.2 Datasets utilitzats

Per poder arribar als objectius detallats anteriorment, caldrà fer ús d'un conjunt de dades que tinguin la informació amb la que després s'hauran d'enfrontar els models. Des del portal “Kaggle”, on es poden trobar molts ”datasets”de qualsevol àmbit, se'n descarreguen uns quants:

- Mental Health Twitter (Twitter 3)[8]
- Depression social media (Twitter Scale) [9]
- Depression: Reddit Dataset (Reddit) [10]

S'agafa més d'un perquè tots tenen alguna característica especial que els fa únics. El primer és el més senzill, doncs únicament té 3 atributs i dues classes. El segon es caracteritza per estar classificat en una escala de valors que defineixen el nivell de depressió, a diferència de la resta on únicament tenen valors de “depressió” o “no depressió”. El tercer es caracteritza per estar ja netejat d'emojis, links d'Internet i similars. Cal comentar que els tres ”datasets” estan desbalancejats, sent la classe important la que té menys mostres. Per aquesta raó s'utilitza el ”macro average”a les mètriques per donar més importància a la classe minoritària.

3.3 Preprocessament inicial

Prèviament a començar a entrenar els models, cal netejar i fer un preprocessament de les dades. Normalment les dades sense tractar són molt difícils de treballar i contenen informació irrellevant. Pel cas de SA es realitza les següents modificacions:

- Eliminar noms d'usuari
- Eliminar ”Stop Words” (articles, adverbis, preposicions, etc...))
- Eliminar números
- ”Lemmanization”: donar un sol valor a totes les paraules que signifiquen el mateix, per exemple passar els verbs a infinitiu.
- Eliminar símbols de puntuació.

3.4 Aproximació per Shallow Learning

Fins ara només s'ha modificat el contingut del missatge, però segueix tenint la mateixa forma (un seguit de paraules). Pels mètodes de "Shallow learning", queda pendent extreure les característiques dels missatges per tenir un format que els models puguin entendre. Hi ha dos tècniques molt utilitzades: "Bag-of-Words" (BoW)i "Term Frequency-Inverse Document Frequency" (TF-IDF) [18].

El BoW funciona creant una llista de totes les paraules que existeixen als missatges. Per cada missatge, indica el número de vegades que apareix cada paraula de la llista. El següent exemple mostra com funciona:

- Frase1: El cotxe que vaig comprar és gran i vermell.
- Frase2: Vaig comprar un ordinador molt gran i car.

	Cotxe	Comprar	Gran	Vermell	Ordinador	car
Frase1	1	1	1	1	0	0
Frase2	0	1	1	0	1	1

La taula resultant conté la freqüència de cada paraula a cada una de les frases. Cal destacar que aquest sistema té en compte només els número de vegades que apareix una paraula, però no la seqüència ni l'ordre.

L'altre mètode és el TF-IDF. Funciona donant un pes (importància) a cada paraula. Consta de dos parts. La primera és "Term Frequency" (TF), que és el número de vegades que una paraula apareix al text, dividit entre el total de paraules del text. Si el text té 100 paraules i la paraula en qüestió apareix 3 vegades, el TF és 0.03.

$$TF_(x,y) = \frac{N_x}{N_y}$$

$$x = paraula, y = text$$

La segona part és "Inverse Document Frequency" (IDF). Es basa en calcular el logaritme del número de missatges al "dataset" entre el número de missatges on la paraula en qüestió apareix. Si el "dataset" té 100 missatges, i la paraula en qüestió apareix a 3 missatges, el IDF és 1.5

$$IDF_x = \frac{\log(N)}{N_x}$$

$$N = número\ de\ documents\ al\ dataset$$

$$N_x = número\ de\ documents\ que\ contenen\ la\ paraula\ X$$

El valor final és TFIDF(paraula) = TF(paraula)*IDF(praula)

$$TF - IDF_w = TF_(x,y) * IDF_x$$

3.5 Aproximació per Deep Learning

Per poder treballar amb xarxes neuronals, també cal fer un pretractament de la informació que tenim. En aquest cas s'anomena "Word Embedding", consistint en assignar un vector a cada paraula. A diferència del tractament fet anteriorment, aquest vector guarda informació semàntica, el que permet que pugui ser associat a altres vectors (paraules) segons els diferents contexts gramaticals.

Els vectors que es creen són elements que posseeixen 2 característiques: longitud i orientació, i estan ubicats en plans multidimensionals. Els vectors que representen paraules amb significats similars s'ubiquen més a prop entre sí, i el significat de cada paraula ve donat pel seu respectiu entorn.

Un exemple seria el següent (veure figura 2): es té un vector corresponent a la paraula "King". Aquest està associat al vector de la paraula "Man". Així, si es resta el vector "Man" i es suma el vector de "Woman", quedaria el vector "Queen".

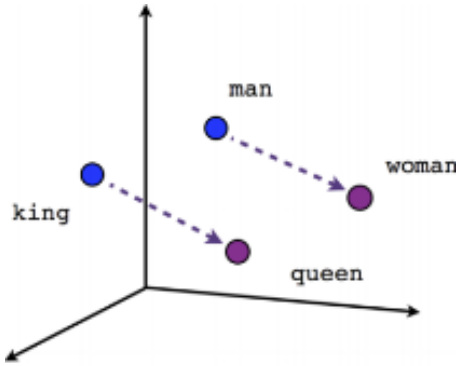

Fig. 2: Exemple de vectors resultants

"GloVe" és una tècnica de "word vectoring" molt potent, i és la utilitzada per aquesta part. A diferència del "Word2vec", una tècnica molt utilitzada en el passat, "GloVe" no depèn només en les estadístiques locals (informació del context local de les paraules), sinó que incorpora estadística global (coocurrencia de paraules) per obtenir els vectors.

3.6 Seguiment de resultats i mètriques

Davant de la quantitat de proves i execucions que cal fer pel treball, cal una manera de guardar-ho tot de forma automatitzada i que permeti accedir-hi fàcilment en un futur. Així doncs, al final de l'execució de cada model s'aprofita per guardar les dades i resultats a un arxiu csv. Les dades que es guarden són model, paràmetres, dataset, type, average type, Recall Score, elapsed time, hyperparameter search i datetime.

La taula 1 mostra el format i informació que té cada camp.

4 RESULTATS

Aquesta secció mostra els diferents resultats obtinguts utilitzant els varis mètodes exposats anteriorment.

4.1 Primers resultats i mètriques útils

Inicialment s'ha fet una execució dels models de "Shallow Learning" per veure com es comporten davant dels "datasets". L'objectiu ara no és veure quin aconsegueix el millor

model	params	dataset	type	average	recall	time	hyperp...	datetime
NB	{"nb_alpha": 45.1}	clean_twitter_scale	tfidf max_features=250	macro	0.53	1.3	True	1/11/2022 16:22
DTC	{"dtc_max_depth...	clean_twitter_scale	tfidf max_features=250	macro	0.53	0.2	True	1/11/2022 16:22
SVM	{"svc_kernel": "...	clean_twitter_scale	tfidf max_features=250	macro	0.72	4.5	True	1/11/2022 16:23

TAULA 1: TAULA DE RESULTATS

resultat, sinó veure el comportament general i la diferència que hi ha entre les mètriques.

La figura 3 mostra els resultats d'aplicar els models al "dataset" de "Reddit".

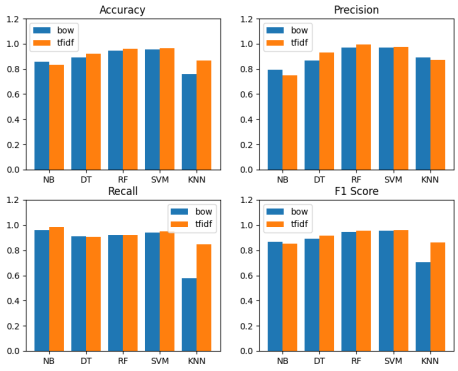

Fig. 3: Resultats dataset "Reddit"

Aquests primers resultats serveixen, a part de veure el rendiment de cada model, per donar més o menys importància a cada una de les mètriques.

Ja que l'objectiu del projecte és poder detectar i analitzar els casos de depressió en xarxes socials. Per tant, és molt més important poder identificar tots els casos reals de depressió. La mètrica "recall" és la que permet saber si l'objectiu s'està complint i, per tant, és la que se li donarà més importància durant el projecte. És cert que no s'estarà tenint molt en compte els casos reals de no depressió classificats com a positius, però al cap i a la fi això són mètodes per passar un filtre inicial, i després cada cas és tractat personalment per persones qualificades (serveis socials, psicòlegs, etc...) És per aquesta raó que és més important fer un filtre inicial que no deixi de banda a cap persona que realment necessiti ajuda, deixant per més endavant separar els classificats positius incorrectament.

4.2 Comparació d'aproximacions del Shallow Learning

Un altre punt a comentar és el paràmetre "max_features" tant del "BoW" com del "TF-IDF". Especifica el màxim número de paraules a extreure'n informació del text. El funcionament normal dels dos mètodes és utilitzar totes les paraules. Clarament, hi ha paraules que apareixen més vegades i aporten més informació que d'altres. Especificant un número es permet indicar el màxim número de paraules a agafar, sent sempre les que més informació aporten.

Els resultats a continuació mostren la diferència de les prediccions utilitzant varis valors pel paràmetre "max_features".

Veient els resultats dels tres datasets, el tipus de model utilitzat és molt més significatiu que no pas el mètode

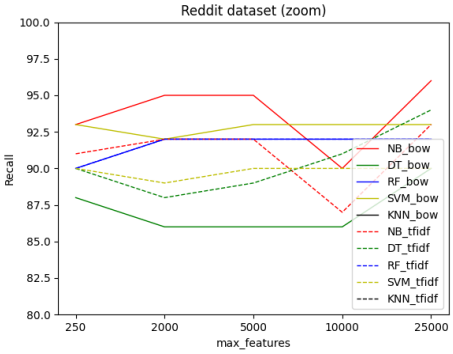

Fig. 4: Resultats dataset "Reddit"

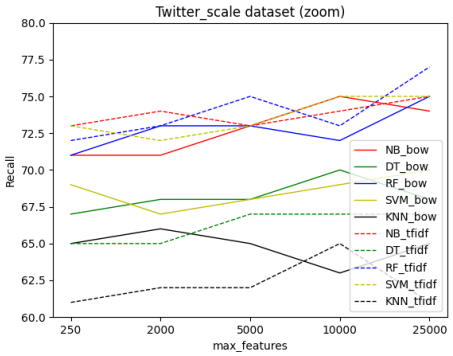

Fig. 5: Resultats dataset "Twitter Scale"

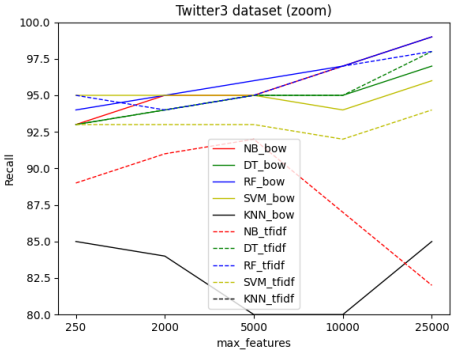

Fig. 6: Resultats dataset "Twitter 3"

per extreure característiques. En canvi, el valor de max_features sí que pot afectar de forma molt significativa al temps d'execució.

No només cal veure el resultats de les mètriques, sinó també el temps d'execució. La figura 7 el mostra tant en escala lineal com logarítmica.

Per una banda es veu que el número de característiques a extreure afecta significativament al temps d'execució. Quan el valor és molt elevant s'utilitzen pràcticament totes les paraules i, depenent de la quantitat de text a analitzar, pot arribar a trigar molts minuts. Utilitzant poques paraules els temps milloren molt, fins al punt que són mil·lèsimes de

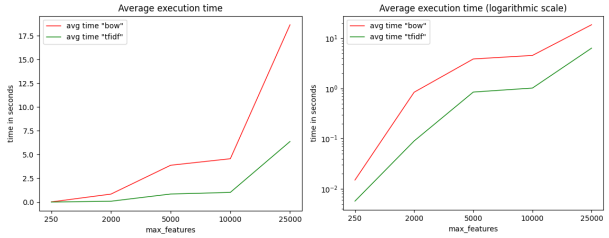

Fig. 7: Temps (escala lineal i logarítmica)

segon. No obstant, ja s'ha vist que els resultats obtinguts són pràcticament idèntics. Per altra banda, el mètode TF-IDF permet tenir resultats entre 2 i 3 vegades més ràpid que el BoW.

El primer gràfic de temps d'execució pot fer pensar que no hi ha gaire diferència de temps entre 250 i 2000 pel valor de "max_features", portant a utilitzar el segon valor ja que dona uns resultats un pel millors. Realment sí que hi ha bastanta diferència entre ambdós valors, podent-se comprovar al segon gràfic, que fa servir una escala logarítmica, que hi ha una diferència de pràcticament 10 vegades més temps entre un i l'altre.

4.3 Confiança de les prediccions

Veient els resultats de les prediccions i els diferents datasets, sembla que el problema ja està resolt en més d'un cas. Les mètriques obtingudes són sorprenentment bones per dos dels tres datasets utilitzats.. Tot i això, cal veure amb quina confiança dona aquests resultats. Al cap i a la fi, s'està buscant la probabilitat de que una mostra pertanyi a cada classe, sent classificada a la classe que tingui la probabilitat més alta.

Ara sí cal analitzar-ho model per model, doncs els resultats són bastant diferents. També es compara utilitzant tant "BoW" com "TF-IDF". Per mostrar els resultats es farà servir el dataset "Reddit", ja que és el que permet veure més bé el comportament de cada mètode. L'apèndix conté els resultats restants dels altres datasets.

El NB (veure figura 8) demostra que realment les prediccions no són gaire confiables. Analitzant primer les prediccions positives, el mètode BoW funciona molt més bé que el TF-IDF, doncs té la gran part de les prediccions amb una probabilitat superior al 80%. De totes formes, les mostres positives tenen en general millors probabilitats que les negatives. Tant per BOW com per TF-IDF la probabilitat de les mostres negatives està distribuïda de forma igual entre el 0% i el 50%, tenint inclús unes quantes mal classificades per sobre del 50

El KNN (veure figura 9), com era d'esperar veient les anteriors gràfiques, té una confiança molt baixa a les prediccions. Aquí es pot veure que el problema el té en els True Positive, doncs les prediccions negatives les fa bé. En canvi, les prediccions positives tenen la probabilitat distribuïda de forma pràcticament igual entre el 0% i el 100%.

El model DT (veure figura 10) té els resultats més

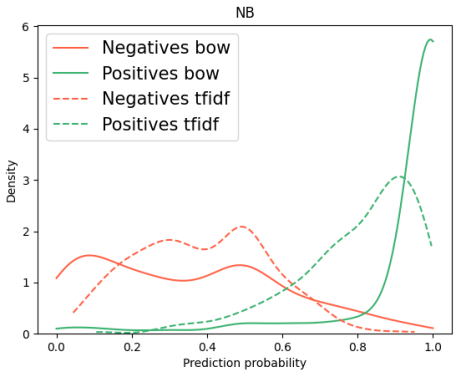

Fig. 8: Confiança Naïve Bayes

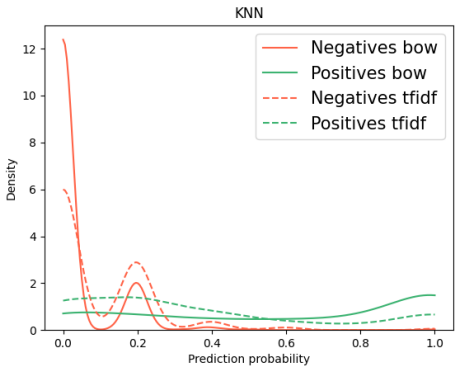

Fig. 9: Confiança K-Nearest Neighbours

confiables fins al moment. Tant per les mostres positives com per les negatives, pràcticament la totalitat d'elles estan correctament classificades amb una probabilitat superior al 85%. És interessant comentar que no hi ha mostres classificades prop del 50% de probabilitat, però sí que hi ha algunes mal classificades a l'altre extrem d'on haurien d'estar.

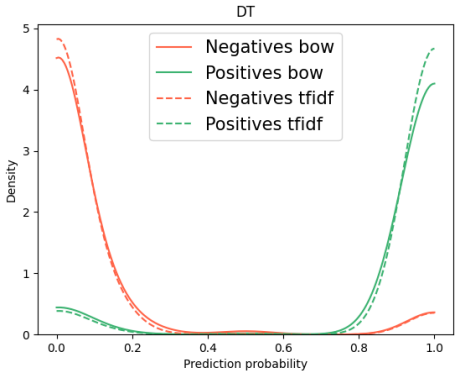

Fig. 10: Confiança Decision Tree

El RF (veure figura 11) és molt similar al DT. Al tractar-se d'un conjunt de DTs, la confiança es queda més centrada ja que es fa una votació d'un seguit de votacions prèvies. Ara ja no es té tantes mostres amb una confiança propera al 100%, però també s'elimina el problema on hi havia mostres mal classificades a l'extrem oposat.

Finalment, els models SVM (veure figura 12) presenten una bona confiança pels dos casos. Si bé no hi ha tantes mostres classificades prop del 100% com altres models,

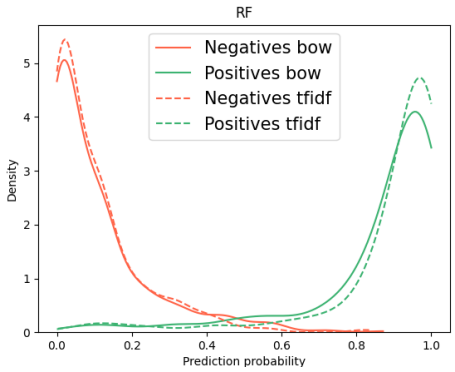

Fig. 11: Confiança Random Forest

segueixen estant a prop i alhora es lliure de tenir mostres mal classificades a l'altre extrem que no els hi pertoca.

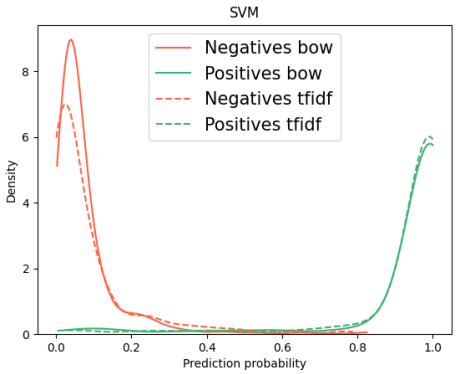

Fig. 12: Confiança SVM

Un cop feta la prova amb 3 "datasets" diferents, es veu que el mètode TF-IDF dona més bones prediccions amb millor confiança, alhora que redueix el temps d'execució a una tercera part del que triga el BoW. L'únic model que té molta diferència entre els dos mètodes d'extracció de característiques és el NB, on TF-IDF dona molta incertesa a les prediccions de classe positiva.

4.4 Hyperparameter Tuning

Mitjançant "Hyperparamter Tuning" s'ha buscat quins són els millors paràmetres per cada model. Addicionalment, es veu si afecta gaire l'elecció d'un valor o altre.

Per realitzar aquesta part es fa servir el framework Optuna [19]. És una eina per Python que permet fer una cerca òptima dels "Hyper Parameters". En comptes de provar valors aleatoris, fa una aproximació basant-se en proves anteriors i poda tots els camins que no són prometedors. Això, juntament amb un ús eficient de paral·lelització, dona uns resultats molt ràpids i òptims.

El resultat de les mètriques no ha tingut cap millora significativa per cap dels models. Tot i això, ha servit per veure quin és el paràmetre que ofereix els millors resultats i quina importància té cadascun d'ells. Les figures a les que es fa referència es poden trobar a l'apèndix.

El SVM, figura 17, té el comportament molt afectat pel "kernel" utilitzat. En les proves realitzades el "rbf" és el que ha donat les millors prediccions.

{"kernel": "rbf", "gamma": 0.73, c: 6.40, "degree": 1}
{"kernel": "rbf", "gamma": 0.47, c: 5.25, "degree": 1}
{"kernel": "rbf", "gamma": 0.80, c: 4.77, "degree": 1}
{"kernel": "rbf", "gamma": 0.54, c: 8.10, "degree": 1}
{"kernel": "rbf", "gamma": 0.80, c: 45.48, "degree": 1}
{"kernel": "rbf", "gamma": 0.80, c: 52.69, "degree": 1}
{"kernel": "rbf", "gamma": 0.52, c: 2.22, "degree": 1}

TAULA 2: MILLORS HYPERPARAMETERS SVM

El KNN, figura 18, sí té més d'un paràmetre que afecta al seu comportament. Com més petit és el número de vens, millor és el resultat. Per altra banda, els millors resultats són tots fent servir la distància de "Minkowski". De totes formes, cal recordar que aquest model dona resultats que queden molt lluny de la resta de mètodes.

{"knn_n_neighbours": 3, "knn_p": 3, "knn_weight": üniform"}
{"knn_n_neighbours": 3, "knn_p": 3, "knn_weight": üniform"}
{"knn_n_neighbours": 4, "knn_p": 3, "knn_weight": "distance"}
{"knn_n_neighbours": 6, "knn_p": 3, "knn_weight": "distance"}
{"knn_n_neighbours": 8, "knn_p": 3, "knn_weight": "distance"}
{"knn_n_neighbours": 5, "knn_p": 3, "knn_weight": "distance"}
{"knn_n_neighbours": 5, "knn_p": 3, "knn_weight": üniform"}

TAULA 3: MILLORS HYPERPARAMETERS KNN

El DT, figura 19, es veu afectat principalment per només un paràmetre, el "dtc_min_samples_leaf". Les proves realitzades que han donat les millors mètriques totes tenen pel paràmetre un valor inferior a 10, i van empitjorant a mesura que el paràmetre augmenta. També cal mencionar que el crieri utilitzada en totes les proves amb bon resultat és l'entropia, mentre que l'índex Gini es queda per sota respecte les mètriques.

{"max_depth": 61, "min_samples_leaf": 9, crit: "entropy", "split": 3}
{"max_depth": 64, "min_samples_leaf": 9, crit: "entropy", "split": 3}
{"max_depth": 60, "min_samples_leaf": 9, crit: "entropy", "split": 3}
{"max_depth": 45, "min_samples_leaf": 1, crit: "entropy", "split": 3}
{"max_depth": 67, "min_samples_leaf": 7, crit: "entropy", "split": 4}
{"max_depth": 47, "min_samples_leaf": 7, crit: "entropy", "split": 4}
{"max_depth": 48, "min_samples_leaf": 9, crit: "entropy", "split": 4}

TAULA 4: MILLORS HYPERPARAMETERS DT

El RF, figura 20, igual que passa amb el Decision Tree, es veu afectat principalment per un sol paràmetre.

{"max_depth": 47, "rf_n_estimators": 23, leaf": 3, "split": 5}
{"max_depth": 28, "rf_n_estimators": 30, leaf": 1, "split": 7}
{"max_depth": 32, "rf_n_estimators": 30, leaf": 1, "split": 8}
{"max_depth": 49, "rf_n_estimators": 19, leaf": 2, "split": 5}
{"max_depth": 28, "rf_n_estimators": 26, leaf": 1, "split": 7}
{"max_depth": 44, "rf_n_estimators": 24, leaf": 2, "split": 6}
{"max_depth": 44, "rf_n_estimators": 23, leaf": 3, "split": 5}

TAULA 5: MILLORS HYPERPARAMETERS RF

El NB, figura 21, té un sol paràmetre. Tot i aparèixer com a millors paràmetres sent un valor molt petit, no és fins que el paràmetre passa del valor 5 que els resultats realment empitjoren. Al cap i a la fi, utilitzar un valor molt alt comporta

portar les probabilitats cap a 0.5 de cada classe i no interessa.

{"nb_alpha": 0.0074749899724023}
{"nb_alpha": 0.0012161141396562502}
{"nb_alpha": 0.0018830314705746475}
{"nb_alpha": 0.005862946278312197}
{"nb_alpha": 0.005247004551393668}
{"nb_alpha": 0.00583714385514146}
{"nb_alpha": 0.005967275794414782}

TAULA 6: MILLORS HYPERPARAMETERS NB

4.5 Comparació de l'aproximació per Deep Learning

Aquesta subsecció mostra i compara els diferents resultats obtinguts mitjançant deep learning.

4.5.1 RNN

A continuació es pot veure unes primeres proves realitzades per veure el rendiment entre una RNN simple i les LSTM i GRU.

Els resultats dels tres datasets presenten un comportament molt similar. La figura 13 mostra els resultats pel dataset "Twitter Scale", la resta es poden trobar a l'apèndix. El tret més destacable és que les RNN simples no són gens útils pel SA pel fet que no poden mantenir el context entre les diferents paraules de la frase. Els altres dos mètodes acaben obtenint resultats pràcticament idèntics un cop s'han executat unes quantes "epochs".

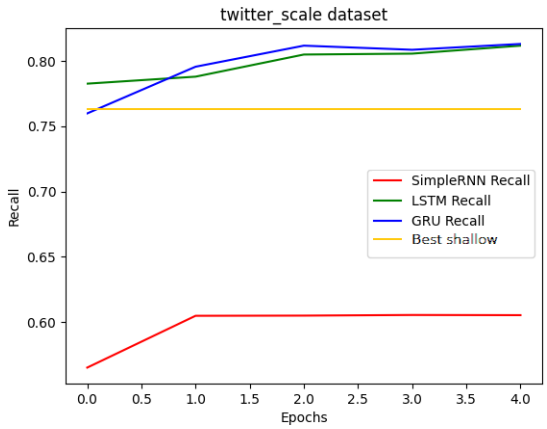

Fig. 13: RNN dataset "Twitter Scale"

No obstant, es pot apreciar que la LSTM sempre està per sobre a la primera "epoch" quan encara no ha tingut cap entrada prèvia ni ha obtingut cap resultat anterior. Aquí és on es mostra com les LSTM funcionen de forma base millor que les GRU pel fet que tenen memòria a curt termini i les ajuda a entendre més bé seqüències molt llargues com poden ser els missatges dels datasets. Fixant-se més bé, als dos datasets de twitter, on hi ha un màxim de 140 caràcters per missatge, a la segona "epoch" LSTM i GRU ja obtenen resultats molt similars. En canvi, el dataset de "Reddit", on els missatges són considerablement més llargs, a la segona "epoch" el LSTM segueix obtenint resultats per sobre de GRU.

Per un altre banda, també cal veure el temps d'execució de cada tipus de RNN. La figura 14 permet veure com les LSTM triguen pràcticament un 50% més que les GRU, per, amb prou feines, oferir millors resultats.

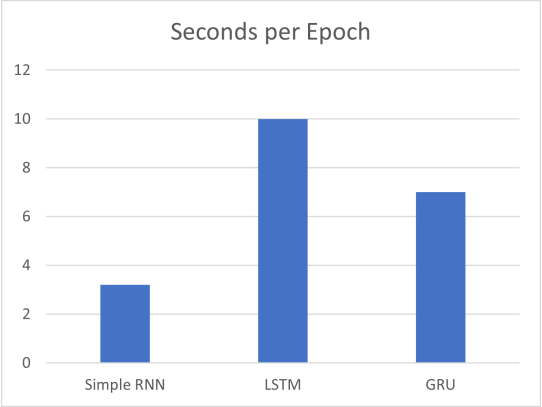

Fig. 14: Temps execució RNN

Un altre aspecte a tenir en compte és com afecta ara el preprocessament de les paraules. Per aquesta part s'ha mantingut la neteja de paraules inventades o noms d'usuari, però s'ha eliminat la lemmanització i treure "stopwords". La figura 15 mostra els resultats.

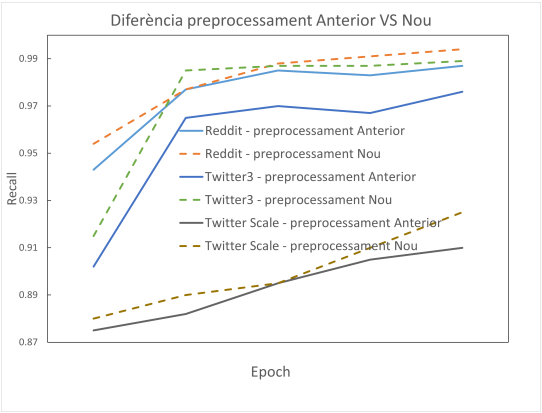

Fig. 15: Diferència al preprocessament amb RNN

Amb el dataset "Reddit"es veu un canvi important respecte els mètodes utilitzats anteriorment. Si amb ells era inviable aplicar-los sense fer un preprocessat molt extrem de les paraules, ara es veu que passa tot el contrari. Si durant el preprocessment es decideix mantenir les "stopwords" i no es fa lemmanització, els resultats obtinguts són lleugerament millors que inclús fent un preprocessat exaustiu. Al dataset "Twitter 3"passa quelcom similar. Es pot apreciar com les xarxes LSTM es poden aprofitar del context de les paraules i inclús aprofitar algunes suposades "stopwords" que realment sí aporten informació al missatge que vol transmetre el text. L'últim dataset, "Twitter Scale", comparteix el mateix comportament que la resta.

Finalment és important fixar-se en el valor de les mètriques. El dataset "Twitter Scale"aconsegueix arribar a un "recall" del 88%. Si es mira els resultats obtinguts amb "Shallow Learning", es queden tots sobre el 75%. És a dir, les LSTM sí que aporten una millora substancial per poder fer bones prediccions.

4.5.2 BERT

Finalment s'ha fet proves amb Bidirectional Encoder Representations from Transformers (BERT), que és el "state-of-the-art" en mètodes de Machine learning per NLP. La figura 16 mostra els resultats pel dataset "Twitter 3". Els resultats pels altres dos datasets es poden trobar a l'apèndix.

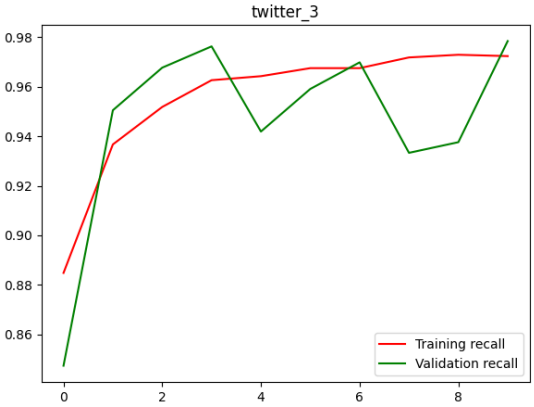

Fig. 16: Resultats BERT twitter3 dataset

Pels tres datasets no es pot apreciar cap millora significativa respecte els models LSTM, amb prou feines incrementa un 1%. El principal inconvenient que tenen respecte les RRN és els recursos que necessiten per ser entrenades, dificultant una millora de rendiment sense haver de dedicar-hi mol més recursos i informació d'entrada.

4.6 Diferències entre les classificacions

Un cop havent fet les classificacions basades en deep learning, s'ha vist que en general són millors que les fetes anteriorment. Mirant amb més detall quins són els canvis, principalment és l'eliminació de falsos positius. A continuació es posen alguns exemples de missatges que ara són ben classificats com a no depressió:

- "study finds no casual relationship between cannabis and depression"
- "dailytonic exposure to the bacteria in soil can be good for mental hearlth and could treat depression and prevent ptsd"
- "just killed two spider and I feel good"
- "don't be sad, armys are here for you we will always suport you btstwt be strong"

Veient el contingut d'aquestes frases, hi ha paraules com "depression", "kill" o "sad" que la majoria de vagades que surten és a les mostres classificades com a depressió. El context de les frases en qüestió, en canvi, deixa clar que no es tracten de missatges depressius. L'ús de les xarxes LSTM o BERT, que permeten tenir guardat la relació que tenen les paraules entre elles i així poder mantenir el significat d'una seqüència de valors concrets, són segurament el que permet aquesta millora.

5 Conclusions

Per una part es pot parlar dels mètodes més tradicionals. En ells s'ha vist que, si bé poden aconseguir prediccions mínimament bones, tenen algunes mancances.

Els classificadors que han donat més bons resultats són els basats en classificació probabilística, és a dir, Naive Bayes (NB) i Random Forest (RF).

Addicionalment, també cal comentar com el preprocessament de les dades afecta significativament el seu rendiment. Primer cal fer una neteja del text, treient paraules que no aporten informació o que no existeixen. Sense fer-ho, els mètodes tradicionals no poden assolir mètriques superiors al 70

Per poder crear una informació que els models puguin entrendre s'ha poast a prova els mètodes Bag of Words (BoW) i TF-IDF. Tant per un cop per l'altre, els resultats tendeixen a millorar lleugerament com més paraules es facin servir durant l'entrenament. El benefici en mètriques, però, és en prou feines d'un 1%, mentre que el temps d'execució puja de forma exponencial per tots. Tot i això, el temps d'execució en TF-IDF és 10 vegades inferior de mitjana i només repercuteix en les prediccions en un 1%. A excepció dels casos on sigui molt important una bona predicció, utilitzar TF-IDF amb un número de paraules d'entre les 5000 i 10000 és el que aporta un millor equilibri entre rendiment i eficiència.

Aquests models també han permès veure la confiança o probabilitat de ser classificat en una classe o altre. Aquí és on es diferencien el Naive Bayes amb el Random Forest. Mentre que el NB té dificultats per donar amb certesa les classificacions negatives i les positives en TF-IDF, el RF té les probabilitats de totes les prediccions positives i negatives prop del 100% i 0% respectivament.

Finalment s'ha intentat fer ús de la cerca d'hiperparàmetres per cada un dels models. La realitat ha sigut que cap d'ells ha tingut cap millora significativa. No es tracta d'un dels casos en que és difícil trobar els paràmetres adequats, sinó que posant els més lògics pel problema ja s'arriba a un bon resultat.

Per una altra banda, s'ha de comentar el paper que tenen els mètodes basats en deep learning. Es veu que fent servir una Recurrent Neural Network (RNN) simple no hi ha suficient per obtenir bones prediccions, és inclús bastant inferior als mètodes anteriors. Utilitzant GRU i LSTM sí que s'aconsegueix millorar molt els resultats gràcies a poder mantenir el context de les paraules i com canvien el significat final del missatge, sobretot en els datasets que anteriorment obtenien les prediccions més baixes.

També s'ha vist que LSTM té més bons resultats quan el missatge és relativament llarg. Mentre que GRU està una mica per sota, LSTM fa ús de memòria a curt termini i permet mantenir el context de les paraules en les seqüències més llargues. És cert, però, que aquesta petita millora es veu afectada per un increment del temps d'execució d'un 50%.

Si bé també es fa un preprocessament de les dades en els mètodes basats en deep learning, ara ja no es fa de forma tant severa. Gràcies a la pròpia naturalesa de GRU i LSTM, es poden deixar les "stopwords" i es prescindeix de fer una lemmanització. Es segueixen esborrant paraules com poden ser urls, noms d'usuari o números. Els resultats fent servir aquest nou tractament són millor que si s'aplica el tractament dels mètodes anteriors.

Finalment, utilitzant BERT no s'ha aconseguit cap millora significativa respecte LSTM o GRU. Amb prou feines incrementa la mètrica en un 1%, però els recursos que cal destinar a l'execució són molt superiors.

De totes formes, tots els mètodes basats en xarxes neuronals han aportat els mateixos canvis a les prediccions. Ara són capaços de solucionar les mostres que anteriorment es classificaven com a falsos positius. En totes aquestes mostres hi ha en comú que apareixen paraules com "depression" o "kill", però que també tenen paraules com "no" que neguen el seu significat. Ara és possible veure quin és el context i la relació que tenen les paraules, mentre que els mètodes anteriors ho passen per alt.

El problema que segueixen tenint tots els mètodes és per identificar els falsos negatius que fan servir ironia o sarcasme. Al cap i a la fi, fan servir unes paraules molt positives o donen un missatge idèntic als negatius reals.

BIBLIOGRAFIA

REFERÈNCIES

[1] Yasar, K. (2022, 12 abril). *social networking.*, from shorturl.at/gCOT7

[2] boyd, D. Social Network Sites *Journal of ComputerMediated Communication, 13*, from shorturl.at/yFSW1

[3] Chan, T.(2020, December 5). *Cyberbullying on social networking sites*& Management, from shorturl.at/iwzEV

[4] Rosenquist, J. N. (2010, 16 marzo). *Social network determinants of depression*, from https://www.nature.com/articles/mp201013

[5] Dey, L. (2016, 31 octubre). *Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier*, from https://arxiv.org/abs/1610.09982

[6] *Sentiment analysis on Twitter data using KNN and SVM - semantic scholar.*, from shorturl.at/fjFPY

[7] *Sentiment analysis on Twitter data using KNN and SVM - semantic scholar.*, from shorturl.at/fjFPY

[8] shorturl.at/stRUV

[9] shorturl.at/rCEH6

[10] shorturl.at/gDGY2

[11] shorturl.at/jN158

[12] Claesen, M. *Hyperparameter search in machine learning.* arXiv.org, from https://arxiv.org/abs/1502.02127

[13] Bernardo, M.(2020, May 18). *Comparing machine learning and deep learning* IRIS., from https://cris.fbk.eu/handle/11582/322156

[14] Vaswani, A (2017, December 6). *Attention is all you need*, from https://arxiv.org/abs/1706.03762

[15] Kokab, S. T *Transformer-based deep learning models*, from shorturl.at/gtzS8

[16] *Overview of the Transformer-based models for NLP tasks.* IEEE Xplore. (n.d.). Retrieved October 6, 2022, from https://ieeexplore.ieee.org/abstract/document/9222960

[17] *LSTM and GRU neural network performance comparison study*, from https://ieeexplore.ieee.org/document/9221727

[18] Pimpalkar, A. P. (2020). *Influence of Preprocessing Strategies on the Performance of ML Classifiers*, from shorturl.at/osvIO

[19] Networks, T. A. P.(2019, July 1). *Optuna: A Next-generation Hyperparameter Optimization Framework*, from https://dl.acm.org/doi/abs/10.1145/3292500.3330701

APÈNDIX

Hyperparameter Search: Importància dels paràmetres

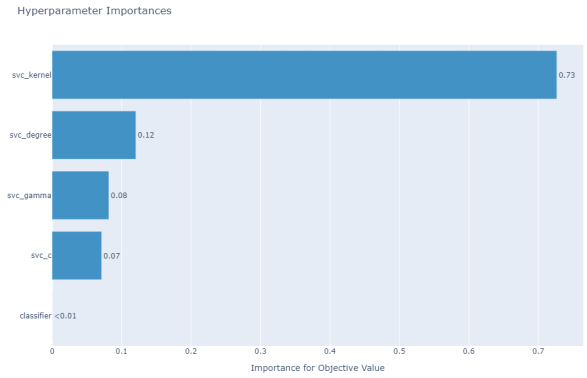

Fig. 17: Importància paràmetres SVM

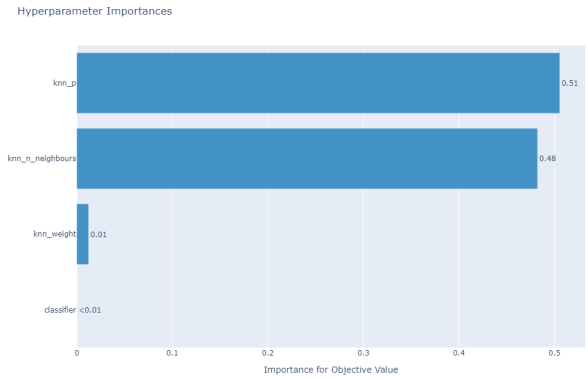

Fig. 18: Importància paràmetres KNN

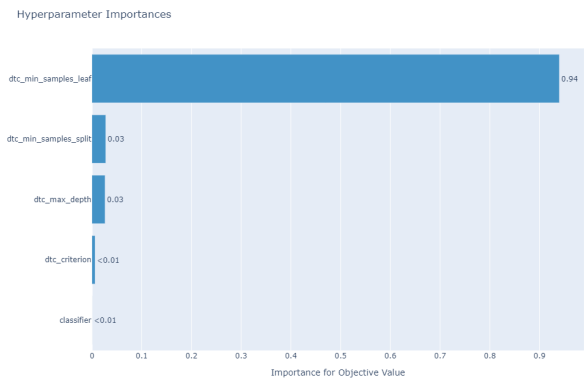

Fig. 19: Importància paràmetres DT

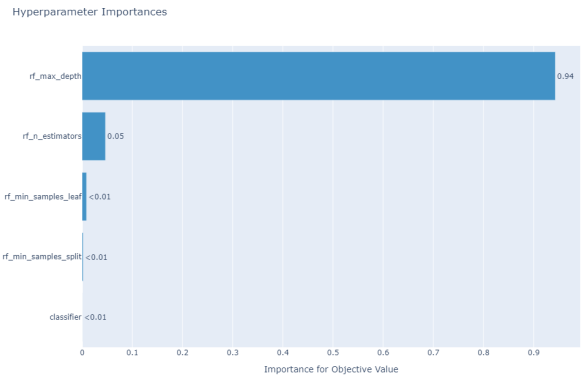

Fig. 20: Importància paràmetres RF

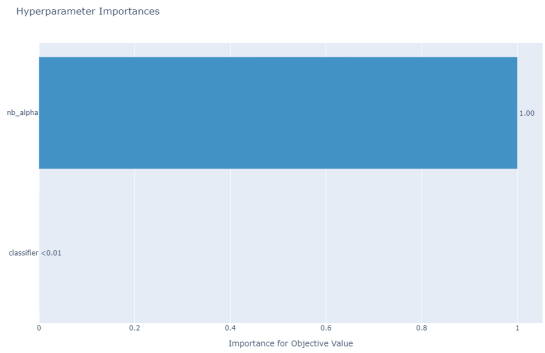

Fig. 21: Importància paràmetres NB

Resultats BERT

Fig. 22: Resultats BERT reddit dataset

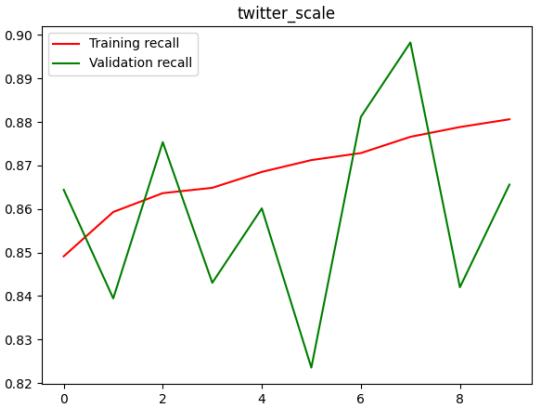

Fig. 23: Resultats BERT twitter Scale dataset

A.1 Diferència RNN

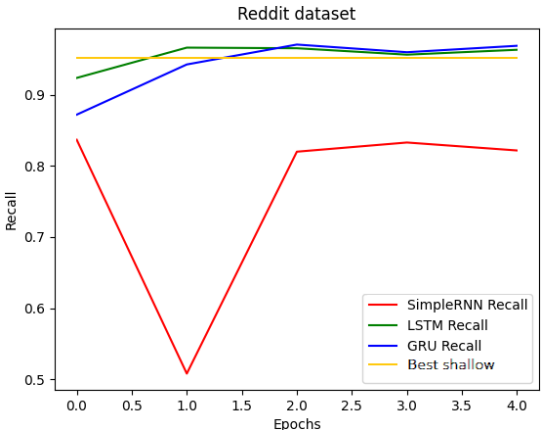

Fig. 24: RNN dataset ”Reddit”

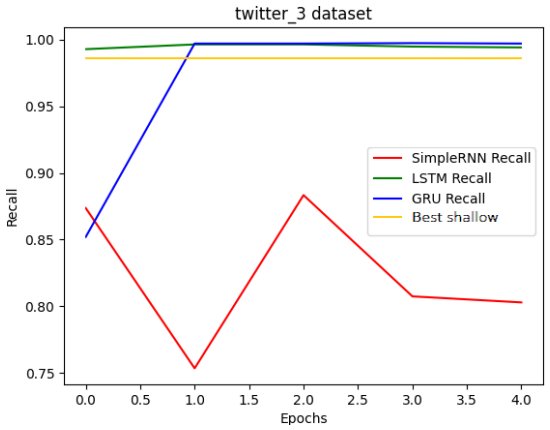

Fig. 25: RNN dataset ”Twitter 3”

A.2 Diagrama de Gantt

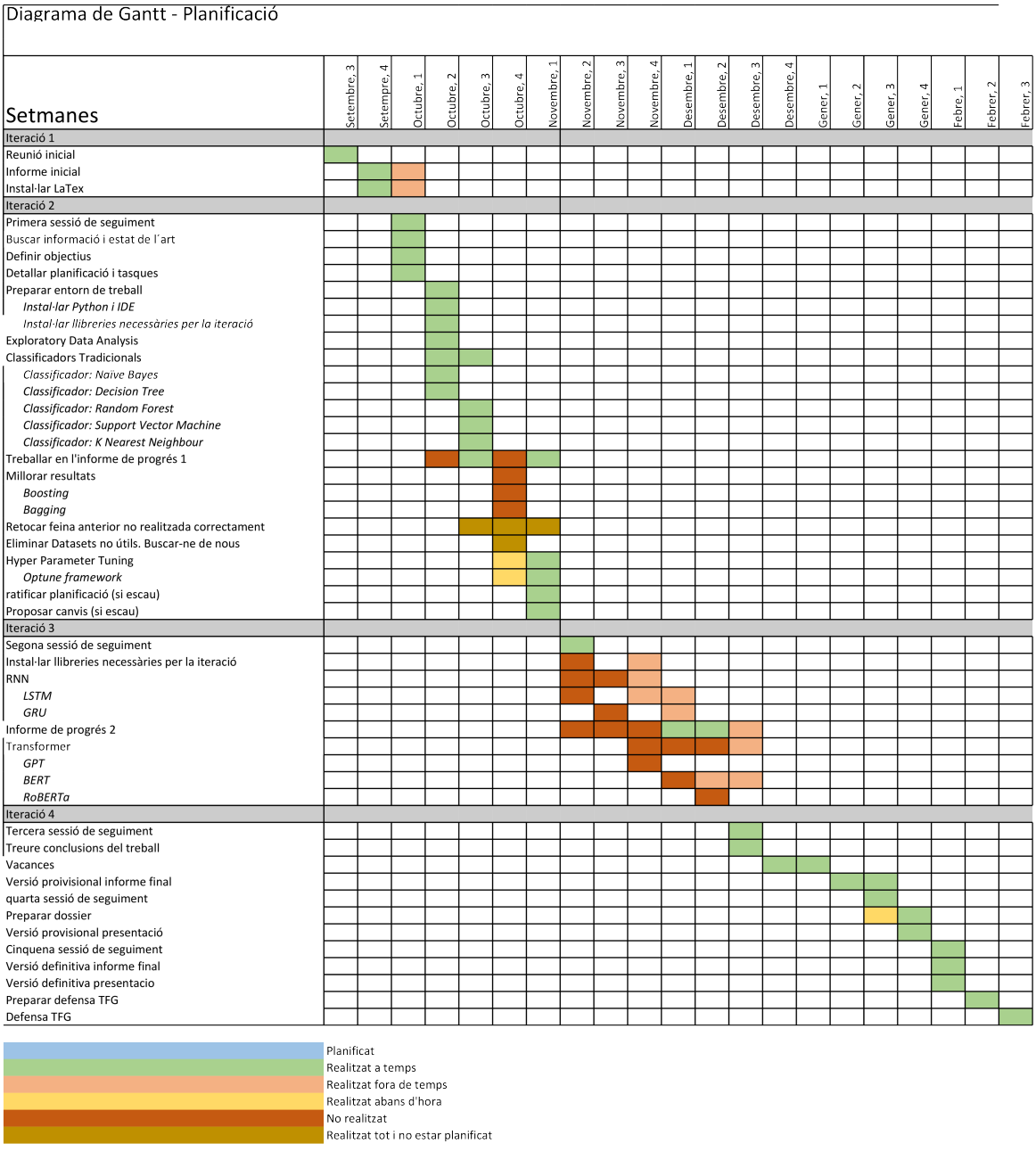

Fig. 26: Evolució del diagrama de Gantt