Facultat d'Informàtica de Barcelona

GIA UPC

Processament del llenguatge humà

Pràctica 2

Alumnes :
Casanovas Poirier, Anna
Pumares Benaiges, Irene

Tutors :
Turmo Borrás, Jordi
Medina Herrera, Salvador

April 12, 2024

Contents

Abstract

L'objectiu d'aquesta pràctica és classificar opinions de pel·lícules per saber si la opinió ha estat positiva o negativa. S'utilitzen diferents mètodes d'aprenentatge supervisat (com el Random Forest o la logístic regression) i també d'aprenentatge no supervisat fent servir llibreries com el SentiWordnet.

Els models són comparats i evaluats mitjançant mesures com la precisió o la matriu de confusió. Finalment, es fa una anàlisi dels resultats obtinguts i es proporcionen unes conclusions sobre el funcionament del model i possibles millores.

1 Introducció

1.1 Motivació del treball

En aquestes últimes sessions de la assignatura, s'ha parlat de la semàntica i de relacions entre paraules. Es va introduir el concepte dels synsets per poder desambiguar el significat d'una paraula. Havent entès aquests conceptes vam descobrir el Wordnet i Sentiwordnet. El darrer serveix per donar-li un valor positiu, negatiu o neutre a cada paraula. Aquestes llibreries ens permeten classificar frases i textos en funció del seu valor positiu o negatiu.

Identificar un text com a positiu i negatiu ens permet saber la opinió de la gent sobre un tema sense haver de llegir o escoltar tota la opinió. En aquesta pràctica farem això mateix. A partir de opinions de pel·lícules, aprendrem a identificar si són opinions positives o negatives.

1.2 Descripció de la tasca i les dades

Les dades que se'ns han proporcionat, les agafem de la llibreria nltk (Natural Language Toolkit- movie_reviews). Comptem amb 1000 resenyes (cadenes de text) positives i mil negatives. L'objectiu de la pràctica és realitzar diferents models d'aprenentatge supervisat i no supervisat per classificar bé aquestes reviews.

Per dur això a terme, per la part supervisada, s'han dividit les dades en train i test de manera que hi ha 700 ressenyes de cada al train i 300 positives i 300 negatives al test.

Per poder fer els models, ha estat necessari aplicar una funció de preprocessament a cada una de les reviews. S'ha aplicat una semblant a la de la pràctica anterior (identificació d'idioma). S'han aplicat els següents canvis:

- S'han eliminat les stopwords
- S'han eliminat els dígits.
- S'ha passat tot el text a minúscules
- Eliminació de caràcters alfanumèrics que no siguin espais.
- Eliminació d'espais adicionals.

A continuació és farà una explicació dels models utilitzats.

2 Models amb aprenentatge supervisat

Per a treballar els models d'aprenentatge supervisat, s'han entrenat: **Logistic Regression**, **Random Forest**, **Naive Bayes** i **XGBoost**.

Per a tots els models s'utilitza CountVectorizer, per a fer el recompte de paraules. La funció ens retorna una matriu on sabrem quantes vegades surt la paraula a les dades. A més, per a logistic regression, random forest i KGBoost es fa ús de Grid Search per a trobar les millors combinacions d'hiperparàmetres, amb 3-fold cross-validation.

2.1 Regressió logística

La regressió logística és un mètode de d'aprenentatge automàtic que serveix per a la classificació binària, és a dir, modela la probabilitat de que una observació pertanyi a una classe particular i utilitza la funció sigmoide per a mapejar la sortida a valors entre 0 i 1 (negatiu i positiu, en aquest cas).

Hi ha diverses raons per a la utilització d'aquest model en aquest cas. En primer lloc, és altament interpretable i eficient computacionalment, el que el fa molt útil permetent entrenar models ràpidament, fins i tot amb conjunt de dades grans. A més a més, és especialment adequat per a problemes de classificació binària, com és el cas, i pot proporcionar probabilitats estimades per a cada classe.

Els hiperparàmetres utilitzats són els següents:

- **max_iter**: controla el nombre màxim d'iteracions per a la convergencia de l'algoritme.
- **c**: aquest paràmetre és l'invers a la força de la regularització. Un valor més petit de c indica una regularització més forta.
- **random_state**: estableix la llavor aleatòria utilitzada per l'algoritme durant l'ajustament del model. Al fixar-la, es garanteix la reproductibilitat dels resusltats.

Es decideix provar totes les combinacions dels següents valors per a cada hiperparàmetre:

max_iter: [100, 300, 600, 1000]
c: [0.001, 0.01, 0.1, 1, 10]
random_state: [42]

Finalment, la millor combinació de valors ha sigut: 100, 0.1 i 42, respectivament.

Després d'entrenar el model, aquests són els resultats:

Classification Report:

	precision	recall	f1-score	support
0	0.82	0.85	0.84	300
1	0.84	0.82	0.83	300
accuracy			0.83	600
macro avg	0.83	0.83	0.83	600
weighted avg	0.83	0.83	0.83	600

Figure 1: Mètriques de la regressió logística

Com es pot observar en les mètriques de rendiment, és un bon model, ja que té un accuracy del 83% i les prediccions de les dos classes estàn bastant balancejades. Els resultats mostren una precisió del 82% i del 84% per a cada una de les classe, el que indica que el model és capaç de predir amb èxit ambdues classes.

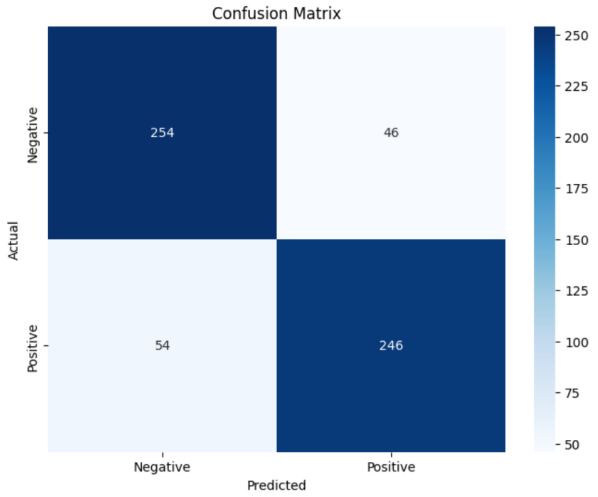

Figure 2: Matriu de confusió de la regressió logística

A la matriu de confusió es veu clarament la classificació de l'algoritme, que és bastant satisfactòtia, tot i que s'equivoca en gairebé un 20% dels casos.

2.2 Random Forest

El Random Forest és un tipus d'algoritme que s'utilitza tant per a problemes de classifiació com de regressió. Es basa en la idea de crear múltiples arbres de decisió durant el procès d'entrenament i combinar les seves prediccions per obtenir una predicció final més robusta i precisa.

L'ús d'aquest algoritme per a classificar ressenyes de pel·lícules ofereix avantatges significatius, ja que és molt robust davant el sobreajustament degut a la combinació dels diversos arbres de decisió.

Els hiperparàmetres utilitzats són els següents:

- **n_estimators**: especifica el nombre d'arbres del bosc. Com més arbres s'utilitzin, més robust serà el mode, però també augmentarà el temps d'entrenament i la complexitat del model.
- **min_samples_split**: nombre mínim de mostres necessàries per a dividir un node intern, és a dir, si el nombre de mostres és menor que aquest valor, el node no es dividirà més.
- **min_samples_leaf**: nombre mínim de mostres necessàries per a que un node sigui considerat com una fulla, és a dir, que no es pot dividir més
- **max_features**: nombre màxim de característique que es consideren per a dividir un node. Les opcions 'log2' i 'sqrt' indiquen que es consideraràn el logaritme en base 2 i l'arrel quadrada del nombre total de característiques, respectivament.
- **max_depth**: profunditat màxima dels arbres de decisió
- **bootstrap**: indica si el mostreig d'arrancada s'utilitza o no per a crear les mostres individuals utilitzades per a cada arbre del bosc.
- **random_state**: controla la reproductibilitat al fixar la llavor aleatòria.

Els valors que utilitzats en el Grid Search per a trobar els millors hiperparàmetres (i els òptims) són els següents:

n_estimators: [5, 10, 20, 50, 80, 100] → 100
min_samples_split: [2, 5, 7, 8, 9, 10, 20] → 10
min_samples_leaf: [1, 2, 4, 8, 9, 10] → 1
max_features: ['log2', 'sqrt'] → 'sqrt'
max_depth: [2, 5, 8, 9, 10] → 10
bootstrap: [True, False] → False
random_state: [42]

Amb aquests hiperparàmetres definits ja es pot entrenar el model i visualitzar els resultats.

Classification Report:

	precision	recall	f1-score	support
0	0.79	0.81	0.80	300
1	0.81	0.78	0.80	300
accuracy			0.80	600
macro avg	0.80	0.80	0.80	600
weighted avg	0.80	0.80	0.80	600

Figure 3: Mètriques del Random Forest

S'observa que el model Random Forest ha aconseguit un rendiment bastant equilibrat també en la classificació de ressenyes de pel·lícules. Tant la precisió com el recall són bastant similars per a les dues classes, indicant que el model no té cap biaix particular cap a cap classe, és a dir, és capaç de predir tant les ressenyes positives com les negatives

amb èxit. L'accuracy del model és del 80%, el que significa que falla en aproximadament un 20% dels casos, això indica un bon renidment general del model en la classificació.

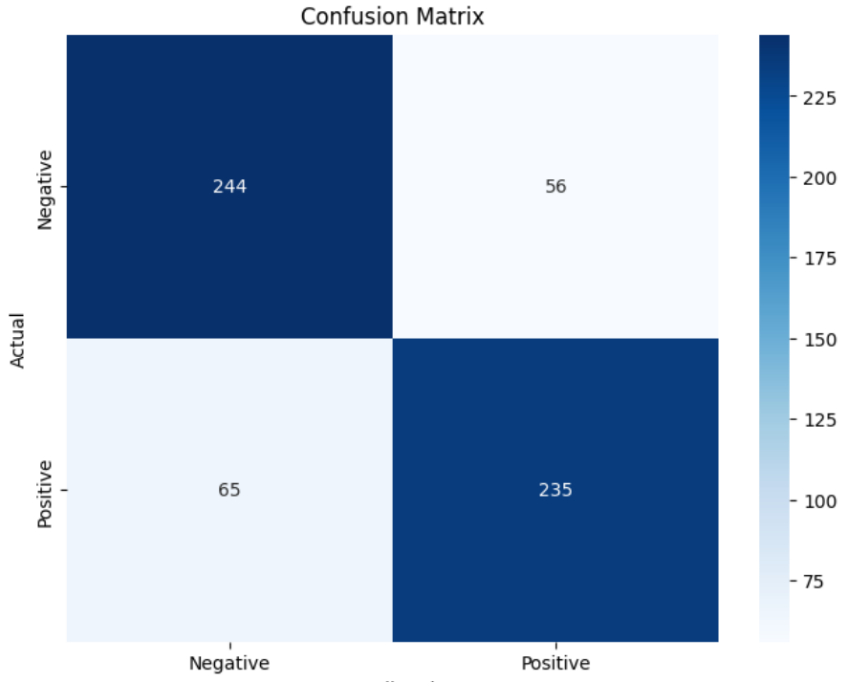

Figure 4: Matriu de confusió del Random Forest

En resum, els resultats són sòlids i equilibrats, com es pot veure en la matriu de confusió, on hi destaquen clarament les observacions encertades, de manera que el model es capaç de realitzar prediccions precises i fiables.

2.3 Naive Bayes

El classificador Naive Bayes Multinomial és un model d'aprenentatge supervisat basat en el teorema de Bayes i està especialment dissenyat per a problemes de classificació amb característiques discretes, com passa amb el processament del text.

És un model bastant eficient computacionalment i pot treballar amb grans quantitats de dades de text de manera ràpida i eficaç. Utilitza la freqüència d'ocurrències de cada paraula en el text com característiques per a realitzar les prediccions.

En aquest cas, no s'ha realitzat cap tècnica per a buscar els millors hiperparàmetres, ja que el model no té hiperparàmetres significatius que necessitin ser optimitzats.

Classification Report:

	precision	recall	f1-score	support
0	0.79	0.81	0.80	300
1	0.81	0.79	0.80	300
accuracy			0.80	600
macro avg	0.80	0.80	0.80	600
weighted avg	0.80	0.80	0.80	600

Figure 5: Mètriques del Naive Bayes

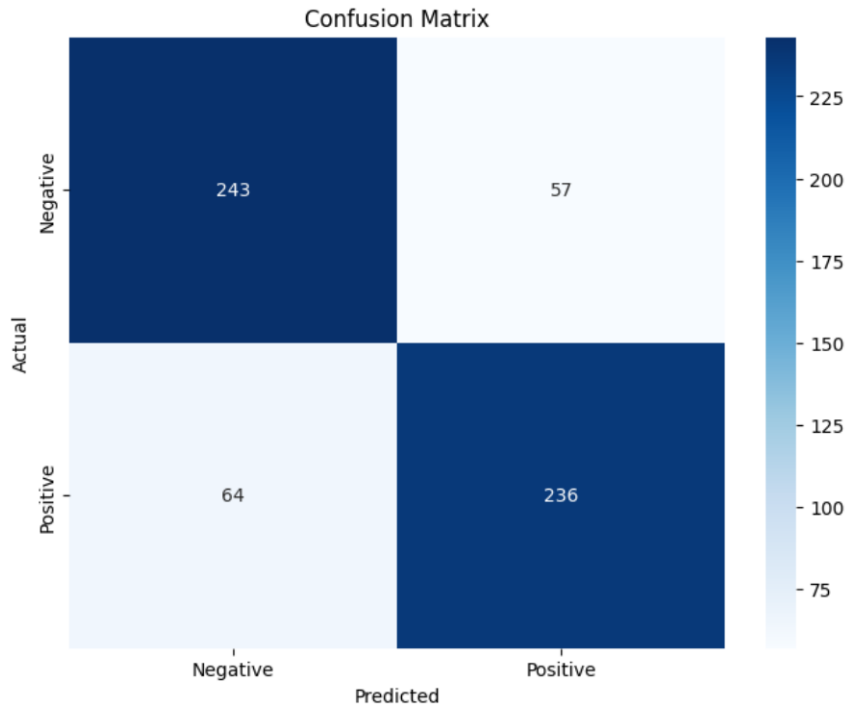

Figure 6: Matriu de confusió del Naive Bayes

En aquest model, també s'observen resultats similars als anteriors, amb un rendiment realativament quilibrat entre la classificació de les positives i negatives.

Els resultats del model suggereixen que té la mateixa precisió que el Random Forest, que és del 80%, per tant, podem dir que els dos són igual d'eficaços per a aquest problema de classificació.

Aquesta observació reforça la robustesa del model Naive Bayes Multinomial en la classificació de text, ja que proporciona un rendiment comparable al d'altres models més complexos com el Random Forest, amb la simplicitat i l'eficiència computacional afegides com a avantatges addicionals. Això implica que el Naive Bayes Multinomial pot ser una opció atractiva en aquest context.

2.4 SVM

Els SVM (Support Vector Machines) són un algorisme de classificació i regressió utilitzat en problemes d'aprenentatge supervisat. Un dels aspectes importants del SVM és la seva

capacitat per trobar l'hiperpla que millor separa les diferents classes, minimitzant l'error de classificació.

La motivació per utilitzar el SVM en aquest treball inclou el fet que aquest algorisme ens va ser ensenyat en el quadrimestre passat. La seva eficàcia i versatilitat fan del SVM una bona elecció per a la nostra anàlisi de ressenyes de pel·lícules.

En aquest cas, l'únic hiperparàmetre que hem utilitzat és el kernel, que defineix el tipus de funció que s'utilitzarà en el model. En aquest cas, hem especificat kernel='linear', que és una opció molt comuna en SVM, i és útil quan es vol un model més interpretable.

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.82	0.82	300
1	0.82	0.80	0.81	300
accuracy			0.81	600
macro avg	0.81	0.81	0.81	600
weighted avg	0.81	0.81	0.81	600

Figure 7: Mètriques del SVM

En aquest cas, també es troben bons resultats, amb una precisió del 81%, una mica millor que els dos anteriors. També s'observa l'equiliibri entre classes confirmant que la predicció no es decanta cap a cap de les dues classes.

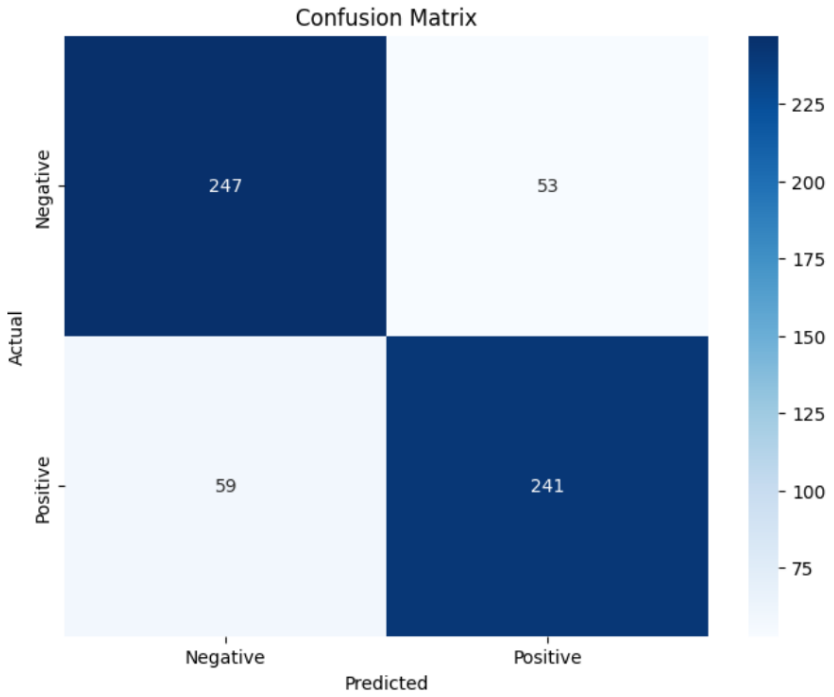

Figure 8: Matriu de confusió del SVM

2.5 XGBoost

XGBoost, o Extreme Gradient Boosting, és un model que utilitza arbres de decisió per a resoldre problemes de classificació i regressió.

La motivació per utilitzar XGBoost en aquest estudi radica en la seva reputació com una de les opcions més eficients i potents disponibles per a la construcció de models predictius.

En aquest cas, si que s'ha fet un Grid Search per a calcular els millors valors dels hiperparàmetres que hem utilitzat, que són els següents:

- **n_estimators**: nomber d'arbres a construir durant l'entrenament. Augmentar aqeust valor pot augmentar la capacita del model, però també el temps i el risc de sobreajustament.
- **max_depth**: profunditat màxima de casa arbre del model.
- **learning_rate**: taxa d'aprenentatge del model
- **gamma**: factor de regualarització que controla la quantitat de penalització per afegir arbres addicionals al model

Es proven diverses combinacions dels hiperparàmetres per als següents valors i finalment els òptims són els següents:

n_estimators: [50, 100, 150, 200] → 200
max_depth: [3, 5, 7, 9] → 3
learning_rate: [0.01, 0.1, 0.3] → 0.1
gamma: [0, 1, 5] → 0

Classification Report:

	precision	recall	f1-score	support
0	0.79	0.82	0.80	300
1	0.81	0.78	0.79	300
accuracy			0.80	600
macro avg	0.80	0.80	0.80	600
weighted avg	0.80	0.80	0.80	600

Figure 9: Mètriques del XGBoost

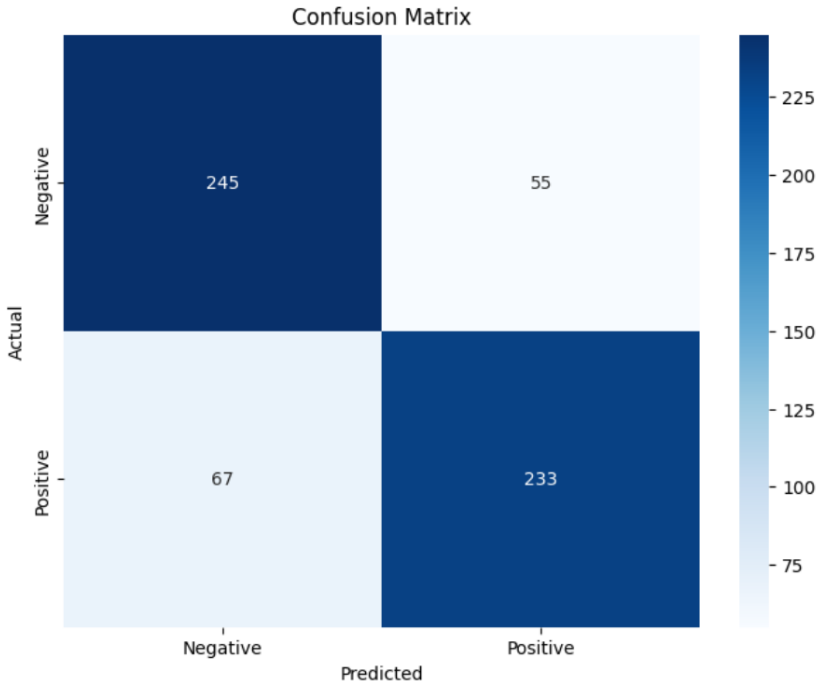

Figure 10: Matriu de confusió del XGBoost

Els resultats del model XGBoost mostren un rendiment també bastant equilibrat en la classifació de les dues classes, com en la resta de models. Tenen una precisió del 80%, com el Random Forest i el Naive Bayes, així que en podem extreure les mateixes conclusions.

2.6 Anàlisi de resultats

En aquesta primera part de models d'aprenentatge supervisat, s'han inclòs: Regressió Logística, Random Forest, Naive Bayes (Multinomial), SVM i XGBoost.

En primer lloc, s'observen les mètriques dels models ordenats de major a menor precisió:

Model	Precisió	Recall	F1-Score
Logistic Regression	0.83	0.83	0.83
SVM	0.81	0.81	0.81
Random Forest	0.80	0.80	0.80
Naive Bayes	0.80	0.80	0.80
XGBoost	0.80	0.80	0.80

Table 1: Comparació de mètriques de rendiment dels diferents models.

Clarament, el millor model, amb una accuracy més gran és el model de regressió logística, amb una precisió del 83%, seguit per l'SVM amb 81% i finalment, els tres restants, que queden empatats amb un 80%.

Model	Precisió		Recall		F1-Score	
	Negatiu	**Positiu**	**Negatiu**	**Positiu**	**Negatiu**	**Positiu**
Logistic Regression	0.82	0.84	0.85	0.82	0.84	0.83
SVM	0.81	0.82	0.82	0.80	0.82	0.81
Random Forest	0.79	0.81	0.81	0.78	0.80	0.80
Naive Bayes	0.79	0.81	0.81	0.79	0.80	0.80
XGBoost	0.79	0.81	0.82	0.78	0.80	0.79

Table 2: Comparació de mètriques per classe dels diferents models.

Si es fa una ullada als valors per a cada classe, s'observa que les ressenyes negatives i positives estan equilibrades en termes de prediccions. Malgrat que hi ha una lleugera tendència a predir una mica millor les ressenyes positives, aquesta diferència no és significativa.

En resum, tots els models models són capaços de fer prediccions relativament precises i fiables per a les dues categories de ressenyes, però si s'hagués de triar un, seria la regressió logística, ja que de tots els models és el que té una precisió general més alta.

3 Models amb aprenentatge no supervisat

En el apartat anterior s'han utilitzat models diferents d'aprenentatge supervisat per poder predir l'etiqueta de cada review com a positiva o negativa. Per fer això, necessitavem partir les dades en un test i en un train, per poder entrenar els models i ajustar els hiperparàmetres i després observar i analitzar els nostres resultats.

En el cas dels models d'aprenentatge no supervisat, ja no fa falta afegir la partició del train, i aplicarem directament el model a les dades del test, perquè l'objectiu d'aquests models no és predir una etiqueta a partir d'una entrada basada en exemples previs d'entrada-sortida. En canvi, els models no supervisats busquen identificar estructures ocultes o patrons dins de les dades sense requerir dades etiquetades per a l'entrenament. Això significa que no necessitem dividir les dades en conjunts d'entrenament i de prova per ajustar paràmetres o fer prediccions basades en etiquetes conegudes prèviament.

Per saber la etiqueta de cada review el que s'ha fet ha estat utilitzar el Sentiwordnet, que ens dirà, per a cada synset, els valors dels sentiments que tenen. Això ens ho dirà per qualsevol tipus de paraula (noms, adjectius, ...).

DESAMBIGUACIÓ DE PARAULES

Una mateixa paraula pot tenir diferents sentits en funció del seu context, això vol dir que pot pertanyer a diferents synsets. El codi intenta trobar el significat (synset) més probable de certes paraules utilitzant l'algoritme lesk. L'objectiu d'aquest algoritme és determinar el sentit d'una paraula en un context, agafant el significat que més comparteix amb les paraules en comú del context (en aquest cas la review especifica actual). Aquesta desambiguació és diferent per cada tipus de paraula possible (depenent de si és un sustantiu, un verb...). És important destacar que tot i que podriem haver fer que el context fós una sola frase, hem fet que sigui tota la review.

Un cop obtinguts els synsets, s'ha fet una búsqueda al SentiwordNet. El codi intenta trobat el sentiSynset corresponent, que ens diu les puntuacions de sentiment associades a aquell synset. Per cada review, inicialment la idea era sumar totes les probabilitats positives i totes les negatives de les paraules i la que donés més alta de les dues seria l'etiqueta escollida.

S'ha de tenir en compte que pot ser que un synset trobat d'una paraula amb el lesk, no trobi el seu corresponent sentisynset. En aquest cas s'ha decidit ignorar la paraula. El codi passa al següent synset sense realitzar cap acció. La paraula no contribuirà a les puntuacions de sentiments acumulades (pos_rating y neg_rating).

TRIA DEL TIPUS DE PARAULES UTILITZADES

El codi que s'ha definit abans es fa només per a certs tipus de paraules. L'objectiu és trobar els synsets de les paraules que creiem que ens donaran informació sobre si una review és negativa o positiva. Hi ha diferents categories de paraules que poden aportar informació sobre el sentiment. Nosaltres hem probat el nostre model amb quatre tipus de

paraules diferents: adjectius, verbs, sustantius i adverbis. A continuació s'explica breument com aquestes classes de paraules poden ajudar a la identificació del sentiment.

- ADJECTIU: és la categoria que a primera vista sembla que està més relacionada, ja que els adjectius descriuren qualitats o característiques dels sustantius. Per descriure una película és important, ja que es poden dir adjectius com aburrit, horrible o excelent, emocionant.

- VERBS: els verbs també ens poden ajudar a trobar la reacció d'un espectador ja que pots utilitzar paraules com estimar, odiar, gaudir o aburrirse. Tot i això, els verbs no sempre aporten aquesta informació aleshores solen ser menys consistent. Hi ha alguns verbs que porten carga emocional clara i altres que la inteció ens dependrà del context.

- NOMS: tot i que solen ser menys expressius que els adjectius i alguns verbs, poden ajudar a reflexar emocions com 'desastre', 'meravella'...

- ADVERBIS: finalment tenim els adverbis, que modifiquen a verbs o adjectius i que poden ser importants per saber el grau d'emoció triada (com per exemple 'terriblement, extremadament'...).

D'aquestes 4 classes a primera vista, sabem que els que ens donaran més informació són els adjectius. Tot i això, els altres també podrien ajudar a la classificació i per tant, amb el conjunt de proba del train que teníem escollit de l'apartat anterior, s'ha mirat amb quina combinació d'aquests dona millors resultats (millor f1-score).

Després de fer una búsqueda de quina combinació de tipus de paraules donaven millor, ens ha sortit que al utilitzar adjectius, verbs i adverbis dona el millor f1-score. Al observar el classification report i veure la matriu de confusió, ràpidament ens hem adonat que hi havia una diferència abismal en l'efectivitat del model per les reviews positives i negatives.

	precision	recall	f1-score	support
Negative	0.75	0.28	0.41	700
Positive	0.56	0.91	0.69	700
accuracy			0.59	1400
macro avg	0.65	0.59	0.55	1400
weighted avg	0.65	0.59	0.55	1400

Figure 11: Classification report amb ['a','v','r']

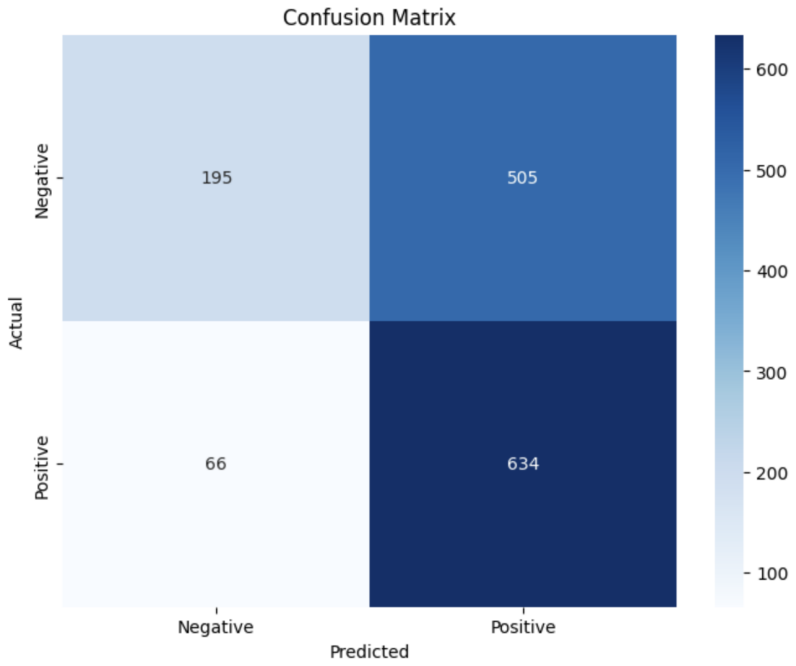

Figure 12: Confusion matrix amb ['a','v','r']

Com es pot veure al classification report, el f1-score de la classe negativa és molt baixa (0.41) i en canvi el de la classe negativa és millor (gairebé 0.7). La matriu de confusió també ens deixa veure que només 200 de 1400 reviews són predites com a negatives i que totes les altres el model diu que són positives. Al reflexionar s'han trobat diverses causes per les quals el nostre model prediu consistentment més reviews positives que negatives.

Primer de tot, pot ser que hi hagi un desequilibri en la manera que s'utilitzen aquests tipus de paraules a les reviews, és possible que els verbs o els adverbis tinguin un biaix cap a ser més positius que negatius (que tinguin més connotacions positives) i aleshores és classifiquin més positives ja que hi ha certs verbs a les review negatives que el Sentiwordnet considera positius.

També pot ser que just els verbs, adverbis i adjectius utilitzats en les reviews positives tendeixin a expressar sentiments més forts (amb un percentatge més alt de positiu). És a dir, que els adjectius positius poden ser més expressius i pertant tenir més influència en les prediccions.

Finalment, el problema podria estar a l'hora de desambiguar les paraules. Si el lesk no ha estat capaç de triar el significat correcte de la paraula, podria ser que es prediguin mostres incorrectament.

Per resoldre el problema, hem analitzat com funcionava el model amb només els adjectius, ja que els verbs i adverbis, com hem comentat, podrien estar fent que es predigui sempre més positiu.

	precision	recall	f1-score	support
Negative	0.67	0.50	0.57	700
Positive	0.60	0.75	0.67	700
accuracy			0.62	1400
macro avg	0.63	0.62	0.62	1400
weighted avg	0.63	0.62	0.62	1400

Figure 13: Classification report amb ['a']

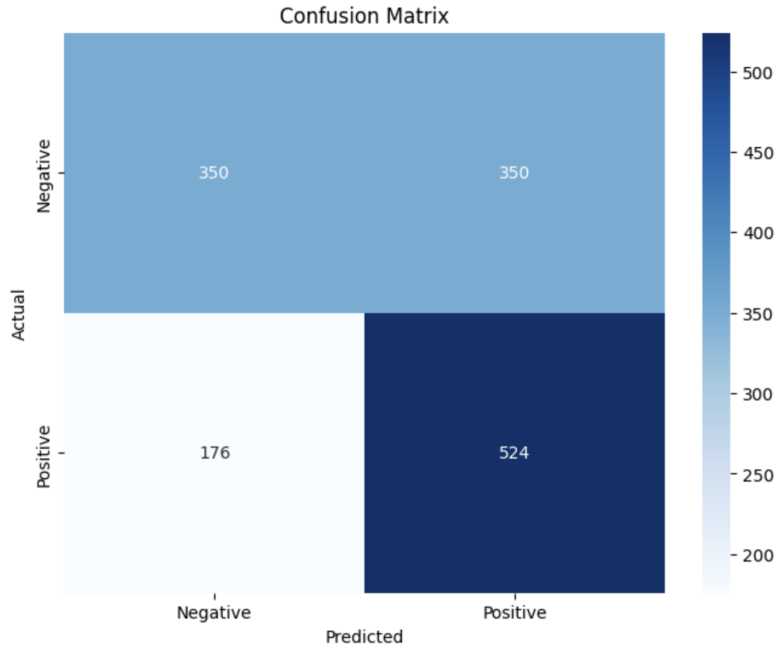

Figure 14: Confusion matrix amb ['a']

A la matriu de confusió i el classification report, veiem que efectivament aquest model està menys esbiaxat cap al positiu. Tot i que prediu més mostres que abans que en realitat són positives com a negatives, prediu molt millor les que de veritat són negatives. Ara ja no hi ha una diferència tan gran entre el f1 score de les dues classes. Com veiem que el f1 score és igual de bo que el anterior, s'ha decidit quedar-nos amb el model que només utilitza els adjectius, ja que com haviem pensat inicialment, són els que han tingut més pes dels tipus de paraules a l'hora de triar la opinió de la review.

Tot i això, encara veiem que hi ha moltes més mostres classificades com a positives que com a negatives. Per això, s'ha pensat afeguir una ponderació. S'ha afegit un valor l que es multiplica amb el negative rating per donarli una mica més d'importància a les paraules que tenen un percentatge negatiu de rating. Aquest valor l inicialment era 1, i ara s'han fet proves variant l entre 1 i 1.5, així donant-li més importància a les paraules que tenen algun percentatge negatiu. Al donar-li el mateix pes a les dues puntuacions presentàvem alguns problemes:

1. Les paraules al dataset pot ser que tinguin puntuacions positives més altes que negatives. Podria ser que per les paraules rellevants a les reviews de películes, la base de dades de sentiments basada en Wordnet tingui puntuacions positives més altes que negatives.

2. També pot ajudar aquest desbalanceig l'algoritme Lesk, ja que per desambiguar paraules, al escollir el sentit que té la definició més semblant al context, pot ser que el context de les reseñes tendeixi a ser més positiu i això podria inclinar la balanza. Això vol dir que si el Lesk no sap desambiguar correctament les paraules, les paraules que poden tenir connotacions tant com negativa com positiva podrien donar-se per positives.

3. També s'ha de tenir en compte que les paraules que són neutres actualment les estem afegint a la suma. Aquestes paraules poden ser les que no tenen un sentiment clar i no haurien de contribuir significativament al balanç dels sentiments. Quan a una paraula que és neutra, igual té un percentatge més gran de contextos positius a la resenya, al no haber un impacte negatiu pot aportar a que les positives predominin.

Per totes les raons anteriors, s'ha decidit afegir una ponderació l per les puntuacions negatives. Això fa que, tot i que no millora extremadament el accuracy i f1 score total, sí que aconsegueix balancejar els resultats perqué el model no predigui sempre com a positiu.

Finalment, hem triat un valor l de 1.13. Ara observem el classification report i la matriu de confusió final amb la ponderació:

	precision	recall	f1-score	support
Negative	0.62	0.55	0.58	300
Positive	0.59	0.66	0.62	300
accuracy			0.60	600
macro avg	0.60	0.60	0.60	600
weighted avg	0.60	0.60	0.60	600

Figure 15: Classification report amb ['a'] i ponderació

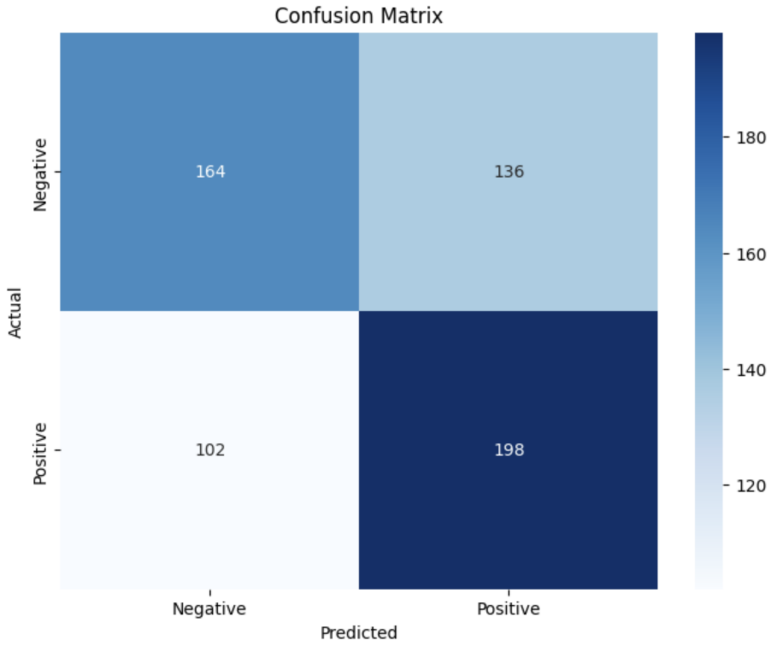

Figure 16: Confusion matrix amb ['a'] i ponderació

El nostre model final, que el fem amb les dades del test que teniem de la partició feta a la primera part de la pràctica, té un f1-score total del 60%. Veiem a la matriu de confusió que en aquest cas es prediuen bé un nombre molt semblant d'opinions negatives i positives (hi ha una diferència com de trenta mostres). Hem aconseguit un model que quan troba un adjectiu amb negativitat li doni importància. És important notar que aplicant-li una ponderació als models que utilitzaven altres categories també millorava però continuaven tenint pitjors resultats.

4 Comparació i anàlisi de resultats i conclusions

Un cop hem fet tots els models supervisats i no supervisats i hem analitzat els seus resultats, podem extreure conclusions i fer una comparació de com de bons són els models.

Primer de tot, queda molt clar que han sortit molt millors resultats als models d'aprenentatge supervisat (on la majoria tenien mètriques de més del 80%). En canvi, els models que s'han fet amb aprenenetatge no supervisat han donat com a màxim 0.6.

Veiem que han classificat millor els models d'aprenentatge supervisat ja que se'ls hi proporciona un conjunt de dades etiquetades prèviament, i aquesta informació ajuda molt a que el model aprengui patrons i relacions entre les característiques de les revisions i les etiquetes. En el altre cas, s'han d'inferir les estructures de dades sense informació prèvia. A més, no tenen feedback explícit sobre el seu rendiment i no poden ajustar i millorar la seva capacitat de classificar. També podriem dir que els algorismes que hem triat per a dur a terme la classificació supervisada (els SVM, Random Forests...) estan dissenyats específicament per la classificació. Finalment, la tria de les característiques important per fer el model és més fàcil amb l'ajuda d'etiquetes.

Tot i això, podríem dir que els algoritmes d'aprenentatge supervisat requereixen de la tria d'hiperparàmetres que és bastant lenta, i això ha fet que aquests triguin molt més en executar-se.

Els errors de classificació en el models supervisats poden tenir diverses causes. Una d'elles podria ser la presència de dades mal etiquetades al conjunt d'entrenament que podrien conduir a prediccions errònies. A més, altres factors com la falta de regularització o la selecció inadequada dels hiperparàmetres poden contribuir a una classificació deficient del model. No obstant això, també cal tenir en compte que no es busca aconseguir una accuracy del 100%, ja que això significaria que el model s'està sobreajustant a les dades d'entrenament i, per tant, no seria capaç de generalitzar bé a dades noves.

Per anar concluint, podem dir que el millor model de tots ha estat la Regressió Logística, tot i que hem aconseguit molt bons resultats en tots els models. El preprocessament de les dades i l'ús del CountVectorizer ha estat de molta ajuda i importància pel funcionament dels nostres models.

En el cas del no supervisat, ens ha ajudat a millorar el seu rendiment triar un tipus de paraules específic (en el nostre cas els adjectius, ja que hem trobat que eren els més indicatius del sentiment d'una review.) i també ens ha ajudat afegir una ponderació.

El percentatge de dades que es classifiquen incorrectament, al voltant del 40%, podria ser degut a la naturalesa del llenguatge, on el significat d'adjectius, adverbis o verbs pot variar en funció del context de la frase on es troben. Això vol dir que paraules que habitualment es considerarien més "positives" poden adoptar un sentit negatiu en altres contextos, i viceversa. Aquesta variabilitat pot causar confusió als models i poden interpretar de manera errònia el significat. A més, altres factors com la ironia o el sarcasme també poden contribuir a la dificultat de la classificació correcta. Per tant, l'error en la

classificació pot ser una conseqüència de la complexitat i la riquesa del llenguatge humà, que sovint no es pot capturar totalment amb els models computacionals.

Aquesta pràctica ens ha ajudat extensament a comprendre millor el material teòric presentat a l'assignatura. Ens ha ajudat a veure desde un punt de vista pràctic l'aplicació del que s'ha estudiat. Hem pogut posar en pràctica tots els nostres coneixements sobre models d'aprenentatge supervisat, i hem creat un model de no supervisat que ha estat una experiència nova.