IT Academy - Projecte Final Data Science - Edaimon De Juan

Construcció d'un Classificador de Sentiments en Ressenyes de Pel·lícules amb Scikit-Learn

Títol

Construcció d'un Classificador de Sentiments en Ressenyes de Pel·lícules amb Scikit-Learn

Objectiu

Aquest projecte té com a objectiu principal desenvolupar un model de machine learning amb la llibreria scikit-learn a Python per predir si una ressenya de pel·lícula és positiva o negativa. S'utilitza un conjunt de dades de l'IMDB amb 50.000 mostres per a l'entrenament i la validació del model.

Dataset

El conjunt de dades prové de l'IMDB i conté 50.000 ressenyes de pel·lícules amb dues columnes: una per la ressenya i una altra pel sentiment associat. L'objectiu és determinar si una ressenya és positiva o negativa.

https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

Metodologia

La metodologia ha estat elaborada per optimitzar la construcció d'un classificador de sentiments en ressenyes de pel·lícules mitjançant tècniques de la llibreria Scikit Learn.

Llibreria scikit-learn
La llibreria scikit-learn, també coneguda com sklearn, és una llibreria de programari lliure d'aprenentatge automàtic per a Python. Aquesta llibreria ofereix diversos algoritmes de classificació, regressió i agrupació, incloent màquines de vectors de suport, boscos aleatoris, impulsió de gradient, k-means i DBSCAN. En l'àmbit d'aquest treball, s'ha optat pels algoritme de classificació, d'aprenentatge supervisat i eines per a la preparació de dades i avaluació de models.

En el context d'aquest treball, s'ha triat l'aprenentatge supervisat com a mètode principal, i per tant s'ha descartat l'aprenentatge no supervisat. Aquesta decisió es basa en el fet que l'aprenentatge supervisat utilitza conjunts de dades etiquetats, és a dir, les dades ja contenen les respostes desitjades. Això és fonamental en tasques com la classificació de ressenyes, on es vol predir la polaritat (positiva o negativa) d'una ressenya. Les ressenyes són valors discrets *per se*, negatius/positius, i aquest és l'origen de la idoneitat d'aquesta llibreria.

```
from imblearn.under_sampling import RandomUnderSampler
from sklearn.model_selection import train_test_split

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC

from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.linear_model import LogisticRegression

from sklearn.metrics import f1_score
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix

from sklearn.model_selection import GridSearchCV
```

Fig. 1: Algoritmes i eines de sklearn utilitzats a aquest treball.
Font: Elaboració pròpia

Les fases en la que es divideix aquest projecte són les següents:

1. Preparació de les Dades

La primera etapa, la "Preparació de les Dades", estableix les bases del procés. Utilitzant la llibreria Pandas, es llegeix el conjunt de dades. Després, es divideixen les dades per crear un desbalanceig entre les classes, per dos raons principals: una, per manipular de manera optimitzada una mostra representativa més petita que l'original de 50.000 registres i, dos, fer servir tècniques de balanceig i de *undersampling*.

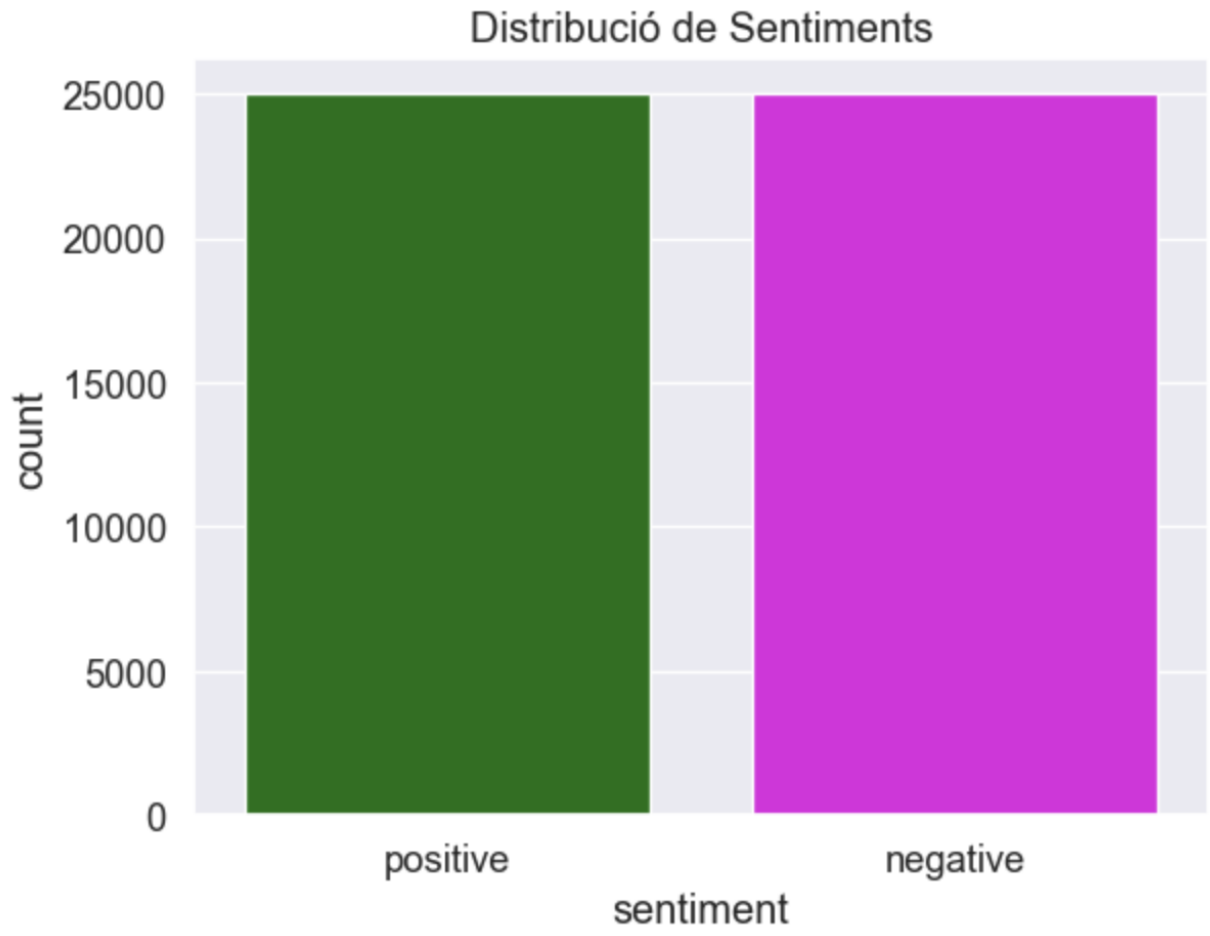

Fig. 2: Distribució de les classes al dataset original.
Font: Elaboració pròpia

Així, es crea una mostra més petita que consisteix en 10.000 files, amb 9.000 ressenyes positives i 1.000 negatives. Aquesta tasca es realitza amb l'ajuda de la tècnica RandomUnderSampler, que permet obtenir un conjunt de dades equilibrat, una característica que es volia afegir per a l'aprenentatge del model d'aquest Classificador de Sentiments.

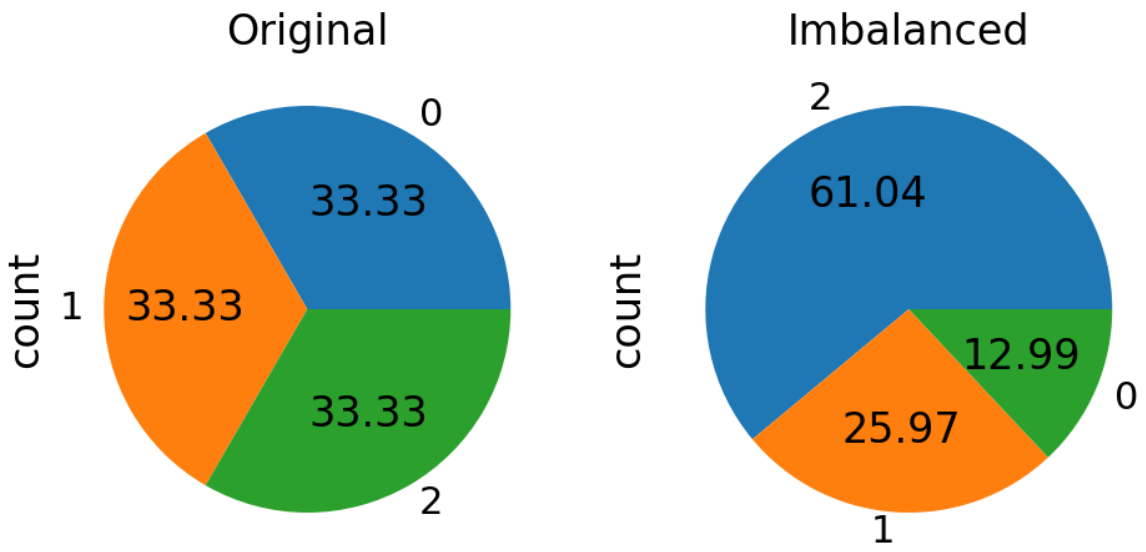

Fig. 3: Representació de la distribució de dades en la classe `imblearn.under_sampling.RandomUnderSampler`
Font: https://imbalanced-learn.org/stable/auto_examples/api/plot_sampling_strategy_usage.html#sphx-glr-auto-examples-api-plot-sampling-strategy-usage-py

2. Representació de Text (Bag of Words):

La següent fase és la "Representació de Text (Bag of Words)". En aquesta etapa, les ressenyes de pel·lícules es converteixen en vectors numèrics mitjançant la tècnica de Term Frequency-Inverse Document Frequency (TF-IDF). Aquest procés té com a objectiu principal transformar les paraules en característiques quantificables, creant així una representació vectorial dels textos. La visualització d'aquestes representacions es realitza mitjançant una matriu dispersa, destacant

les paraules més representatives que contribueixen significativament a la classificació.

La tècnica de "Bag of Words" converteix les ressenyes de pel·lícules en vectors numèrics mitjançant la tècnica de Term Frequency-Inverse Document Frequency (TF-IDF). El "Bag of Words" és un model crític en el processament del llenguatge natural (NLP) i en l'aprenentatge automàtic, ja que representa una col·lecció no ordenada de paraules i les seves freqüències en un document.

3. Selecció del Model:

La "Selecció del Model" constitueix un pas crucial en la implementació del sistema. Diversos models de classificació supervisada, com Support Vector Machines (SVM), Arbres de Decisió, Naive Bayes i Regressió Logística, són explorats per determinar el més adequat per a aquesta tasca. L'avaluació dels models es basa en diverses mètriques, com la precisió mitjana, F1 Score, informe de classificació i matriu de confusió, proporcionant una visió completa del rendiment de cada model.

En més detall, les eines utilitzades han estat les següents:

- Support Vector Machines (SVM) és un algorisme que busca la millor manera de separar dos grups de dades en un espai de característiques, fins i tot si això significa utilitzar una línia imaginària en més dimensions per fer-ho de la manera més precisa possible.

- Els Arbres de Decisió són com un procés de presa de decisions, on es fan preguntes sobre les característiques de les dades per arribar a una decisió final sobre a quin grup pertanyen. Aquest procés es realitza de manera iterativa, dividint les dades en subgrups més petits fins a arribar a una decisió.

- Naive Bayes és un algorisme que es basa en el Teorema de Bayes i assumeix independència condicional entre les característiques per

classificar les dades. Encara que aquesta suposició sigui "ingènua", sovint és eficaç en tasques de classificació de text i altres àmbits.

- La Regressió Logística és un algorisme utilitzat per a la classificació, que utilitza la funció logística per predir la probabilitat que una observació pertanyi a una classe en funció de les seves característiques. Aquesta probabilitat es converteix en una decisió de classe mitjançant un llindar fixat.

```
# Avalua el rendiment del model SVM
svc_score = svc.score(test_x_tfidf, test_y)
print(f"Rendiment del model SVM: {round(svc_score, 2)}")

# Avalua el rendiment del model Decision Tree
dec_tree_score = dec_tree.score(test_x_tfidf, test_y)
print(f"Rendiment del model Decision Tree: {round(dec_tree_score, 2)}")

# Avalua el rendiment del model Naive Bayes
gnb_score = gnb.score(test_x_tfidf.toarray(), test_y)
print(f"Rendiment del model Naive Bayes: {round(gnb_score, 2)}")

# Avalua el rendiment del model Logistic Regression
log_reg_score = log_reg.score(test_x_tfidf, test_y)
print(f"Rendiment del model Logistic Regression: {log_reg_score}")

```

```
Rendiment del model SVM: 0.85
Rendiment del model Decision Tree: 0.51
Rendiment del model Naive Bayes: 0.63
Rendiment del model Logistic Regression: 0.85
```

Fig 4: Resultats del rendiment del model analitzat a través de diferents eines d'avaluació: Support Vector Machines (SVM), Arbres de Decisió, Naive Bayes i Regressió Logística
Font: Elaboració pròpia

Segons el resultat en el treball, s'ha fet una avaluació de rendiment del model amb SVM, amb les següents tècniques d'avaluació:

- F1 Score:
  Aquesta mètrica proporciona un equilibri entre la precisió i la sensibilitat del model. És particularment útil en tasques de classificació on hi ha un desequilibri entre les classes.
- Classification Report:

Aquest informe proporciona una visió detallada de diverses mètriques d'avaluació del model, com ara precisió, sensibilitat, F1-score i suport per a cada classe.

- Confusion Matrix:
  Aquesta matriu ofereix una visió detallada de les prediccions del model, incloent els valors de vertader positiu, fals positiu, vertader negatiu i fals negatiu. Això és útil per entendre com el model classifica les mostres i identificar possibles errors de classificació.

```
# Imprimeix el informe de classificació amb SVC
print("Informe de classificació amb SVC:")
print(classification_report(test_y,
                            svc.predict(test_x_vector),
                            labels=['positive', 'negative']))

```

```
Informe de classificació amb SVC:
              precision    recall  f1-score   support

    positive       0.84      0.87      0.85       335
    negative       0.86      0.83      0.84       325

    accuracy                           0.85       660
   macro avg       0.85      0.85      0.85       660
weighted avg       0.85      0.85      0.85       660
```

Fig 5: Resultats del rendiment de l'informe de classificació amb SVC
Font: Elaboració pròpia

4. Afinació del Model

La fase final, "Afinació del Model", es centra en afinar el rendiment del model SVM. Aquest ajust es realitza mitjançant la tècnica de GridSearchCV, que implica una cerca exhaustiva de paràmetres específics per trobar els valors òptims dels hiperparàmetres del model. Aquest procés de sintonització és essencial per

millorar la capacitat predictiva del model, ja que permet ajustar-lo a les característiques particulars del conjunt de dades i millorar-ne la capacitat de generalització.

Després de completar aquestes etapes, s'ha dut a terme una avaluació dels resultats obtinguts mitjançant l'anàlisi de mètriques específiques. Aquesta avaluació serveix per determinar la eficàcia del model en la classificació de ressenyes positives i negatives, així com la seva capacitat per gestionar casos límit o ambiguïtats en el conjunt de dades. La interpretació de la matriu de confusió ofereix una comprensió detallada dels falsos positius, falsos negatius, verdaders positius i verdaders negatius, contribuint a la identificació de possibles àrees d'ampliació o millora.

Conclusions:

Un dels reptes d'inici, el desbalanceig de les classes en conjunts de dades, ha estat útil per veure com pot afectar la capacitat dels models d'aprenentatge automàtic per fer prediccions precises. Per explorar aquesta qüestió, s'han aplicat tècniques com l'undersampling. Això ha permès comprovar la capacitat dels models per predir amb precisió les classes minoritàries, evitant així el biaix en les prediccions i millorant el rendiment general del model.

Per altra banda, la divisió aleatòria de les dades en conjunts d'entrenament i proves ha ajudat a evitar l'overfitting del model, és a dir, que el model s'ajusti massa als detalls del conjunt d'entrenament i no sigui capaç de generalitzar correctament en dades noves.

En l'àmbit de l'aprenentatge supervisat, s'ha observat que els algoritmes Support Vector Machines (SVM) i Logistic Regression han presentat els millors resultats, amb una precisió del 85%. Això indica que aquests algoritmes han estat capaços de classificar les dades amb una elevada precisió, la qual cosa és crucial en tasques de classificació on es requereixen prediccions acurades.

A més, s'ha aplicat la tècnica de maximització amb GridSearchCV, la qual ha contribuït a aconseguir una precisió del 85,6%. Això indica que les tècniques d'avaluació i optimització dels models han estat efectives en millorar el rendiment dels models i en assegurar prediccions precises.

Com a reptes i millores futures, s'ha identificat la possibilitat d'aplicar tècniques avançades de processament de llenguatge natural i cross-validation per avaluacions més robustes. Això implica la utilització de mètodes més sofisticats per al tractament de dades textuals i la validació dels models, amb l'objectiu de millorar la capacitat dels models per generalitzar correctament i fer prediccions precises en noves dades.

Bibliografia:

baeldung. 'F-1 Score for Multi-Class Classification | Baeldung on Computer Science', 19 August 2020. https://www.baeldung.com/cs/multi-class-f1-score.

'Classification Report — Yellowbrick v1.5 Documentation'. Accessed 1 December 2023. https://www.scikit-yb.org/en/latest/api/classifier/classification_report.html.

Delua, Julianna. 'Supervised vs. Unsupervised Learning: What's the Difference?' IBM Blog, 12 March 2021. https://www.ibm.com/blog/supervised-vs-unsupervised-learning/www.ibm.com/blog/supervised-vs-unsupervised-learning.

GeeksforGeeks. 'Support Vector Machine (SVM) Algorithm', 20 January 2021. https://www.geeksforgeeks.org/support-vector-machine-algorithm/.

KDnuggets. 'Confusion Matrix, Precision, and Recall Explained'. Accessed 12 December 2023.

https://www.kdnuggets.com/confusion-matrix-precision-and-recall-explained.

'Machine Learning Tutorial: The Naive Bayes Text Classifier'. Accessed 3 December 2023. https://blog.datumbox.com/machine-learning-tutorial-the-naive-bayes-text-classifier/.

Okamura, Scott. 'GridSearchCV for Beginners'. Medium, 30 December 2020. https://towardsdatascience.com/gridsearchcv-for-beginners-db48a90114ee.

ProjectPro. 'Generate Classification Report and Confusion Matrix in Python -'. Accessed 6 December 2023. https://www.projectpro.io/recipes/generate-classification-report-and-confusion-matrix-in-python.

'RandomUnderSampler — Version 0.11.0'. Accessed 1 December 2023. https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html.

Roepke, Brian. 'A Quick Introduction to Bag of Words and TF-IDF'. Data Knows All, 00:00:00-08:00. https://www.dataknowsall.com/bowtfidf.html.

'Scikit-Learn: Machine Learning in Python — Scikit-Learn 1.3.2 Documentation'. Accessed 3 December 2023. https://scikit-learn.org/stable/.

'Sentiment Analysis of IMDB Movie Reviews'. Accessed 1 December 2023. https://kaggle.com/code/lakshmi25npathi/sentiment-analysis-of-imdb-movie-reviews.

'Supervised vs. Unsupervised Learning [Differences & Examples]'. Accessed 5 December 2023. https://www.v7labs.com/blog/supervised-vs-unsupervised-learning, https://www.v7labs.com/blog/supervised-vs-unsupervised-learning.

Team, Great Learning. 'Hyperparameter Tuning with GridSearchCV'. Great Learning Blog: Free Resources what Matters to shape your Career!, 30 May 2023. https://www.mygreatlearning.com/blog/gridsearchcv/.

'Text Classification'. Accessed 12 December 2023. https://lena-voita.github.io/nlp_course/text_classification.html.