Documentație proiect IA

Popescu Mihnea – Grupa 364

1. Modelul Linear SVC

Pentru început, am importat dataset-urile din fișierele train_data.csv și test_data.csv și le-am transformat în DataFrame-uri Pandas folosind pandas.read_csv.

Am convertit fiecare etichetă unică (England, Ireland, Scotland) într-un număr întreg, pentru a le putea folosi în antrenarea modelului.

Pentru împărțirea în date de test și antrenare, am folosit funcția train_test_split din pachetul sklearn.model_selection. Am ales să împart întreg setul de date în 75% date de antrenare și 25% date de testare, pentru a preveni supraantrenarea modelului. Singurul parametru folosit a fost random_state, pentru a mă asigura că output-ul acestei funcții este același în cazul apelării multiple, chiar dacă datele vor fi amestecate (pentru a înlătura posibilitatea apariției unei secvențe repetitive ce poate dăuna antrenarea modelului). Am obținut astfel 31.177 date de antrenare și 10.393 date de testare.

Construcția Bag of Words a fost realizată folosind funcția de preprocesare a datelor CountVectorizer din pachetul sklearn.feature_extraction.text. După împărțirea datelor, acestea au fost folosite direct de CountVectorizer, fără vreun parametru adăugat, deoarece am obținut rezultate mai bune cu preprocesarea built-in a funcției CountVectorizer decât atunci când am folosit o funcție separată în care separam cuvintele. Am ales să nu setez o valoare pentru parametrul max_features, deoarece am obținut o precizie mai bună fără această limită. (Din teste, CountVectorizer folosește de obicei aproximativ 200.000 de caracteristici).

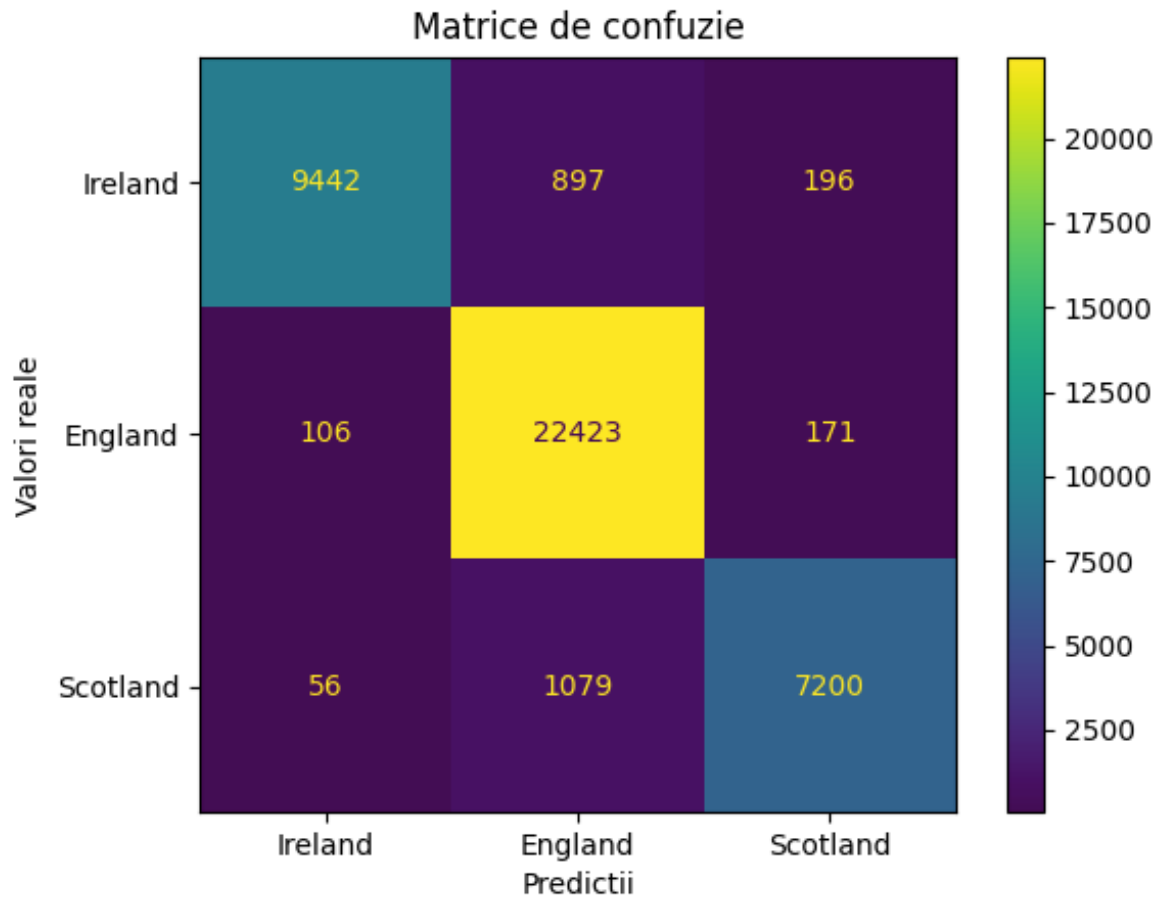
Menționez faptul că am încercat standardizarea datelor cu ajutorul funcției StandardScaler din pachetul sklearn.preprocessing, însă am obținut rezultate mai mici decât cele cu datele nestandardizate.

Tipul de model pe care l-am ales (și cu ajutorul căruia am obținut cele mai bune rezultate) este LinearSVC din pachetul sklearn.svm. Hyperparametrul folosit a fost C (parametrul de regularizare). După multe teste, am ajuns la concluzia că valoarea cea mai bună a C-ului este de 0.005.

Timpul de antrenare a funcției CountVectorizer este de 4.1 secunde, iar antrenarea modelului a durat 33.4 secunde (pe întreg setul de date disponibil).

Rezultatele în urma antrenării în maniera 5-fold cross-validation sunt: [0.29588646 0.47570363 0.55953813 0.38513351 0.52802502], adică o acuratețe medie de 47% cu abaterea standard de 10%.

Matrice de confuzie:



Scorul obținut pe 40% din datele publice de pe Kaggle este: 0.69011

2. Modelul MultinomialNB

Preprocesarea datelor, precum și construcția Bag of Words-ului a fost realizată exact ca în cazul modelului LinearSVC.

În urma testelor, am tras concluzia că cea mai bună soluție a fost oferită prin folosirea unui singur hyperparametru: Alpha = 0.4

Timpul de antrenare a modelului este de 0.09 secunde.

Scorul obținut pe 40% din datele publice de pe Kaggle este: 0.67099