Nume: Spătaru Andreea – Bianca
Master ISI – Anul I

Temă de predicție

– SII –

Pentru detecția știrilor false, folosind acest set de date am început prin citirea celor două documente în scriptul python și de asemenea am definit o listă de stopwords în limba franceză. Am ales să utilizez această listă pentru optimizarea procesului de prelucrare a textului, pentru a reduce dimensiunea datelor și a crește eficiența modelului, făcându-l mai capabil să identifice veridicitatea articolului.

Mai departe, am folosit tehnica TF – IDF (Team Frequency – Inverse Document Frequency) pentru a vectoriza textele articolelor și anume pentru a transforma datele de tip text într-o formă numerică, pe care modelul de învățare automată să o poată procesa.

De asemenea, am ales să utilizez această metodă, deoarece este ușor de implementat, setul de date nefiind extrem de mare și destul de eficientă, îmbunătățind performanța modelului și reduce influența cuvintelor comune, care pot să nu fie semnificative pentru predicție. Pe de altă parte, este o metodă care ajută la captarea semnificației contextuale a cuvintelor, fiind esențial pentru clasificarea corectă a știrilor ca „fake”, „biased” sau „true”.

Pentru a clasifica articolele am folosit ca model regresia logistică, un model de învățare automată de tip "supervizat", folosit pentru rezolvarea problemelor de clasificare binară sau multiclasă.

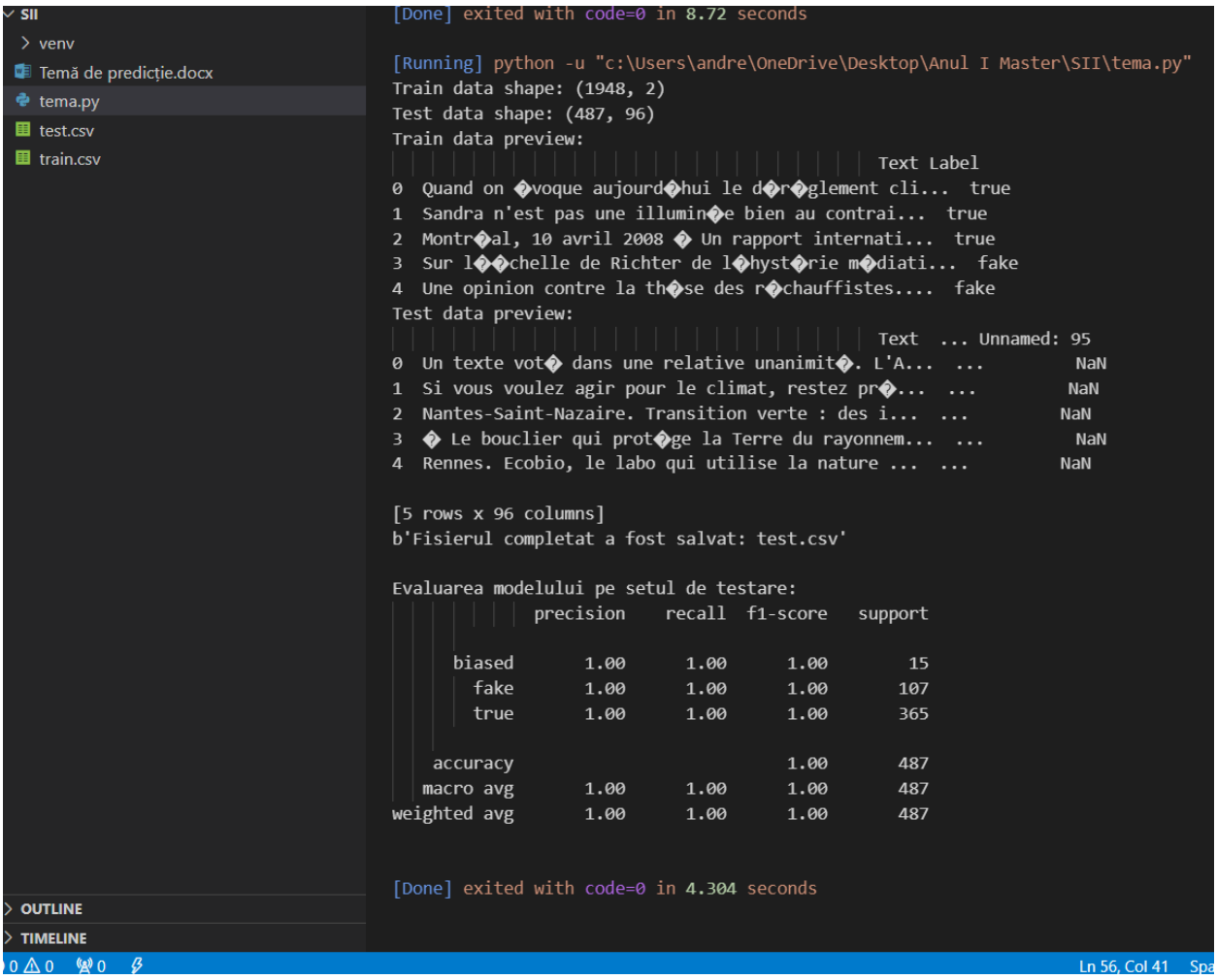
Pentru evaluarea modelului am folosit classification_report din biblioteca scikit-learn, care generează diferite metrici, precum: *accuracy* pentru procentul de predicții corecte față de totalul predicțiilor, *precision* pentru procentajul de predicții corecte pentru fiecare clasă, *recall* pentru procentajul de articole corect identificate din toate articolele relevante pentru fiecare clasă și *F1-score* pentru media armonică a precision-ului si recall-ului, care este utilă atunci când există un dezechilibru între clase.

Nume: Spătaru Andreea – Bianca
Master ISI – Anul I

Iar pentru a evalua acuratețea modelului am comparat etichetele prezise cu etichetele reale ale setului de test.

Modelul a fost antrenat folosind metoda fit(x_train_vec, y_train), unde x_train_vec reprezintă vectorii TF-IDF ai articolelor de știri, iar y_train reprezintă etichetele reale, „fake”, „biased” și „true”.

În concluzie, deși regresia logistică este un model relativ simplu, a funcționat bine pentru acest set de date, iar folosirea TF-IDF a permis vectorizarea eficientă a textului.



Nume: Spătaru Andreea – Bianca
Master ISI – Anul I