FACULTATEA DE AUTOMATICĂ ŞI CALCULATOARE

Splaiul Independentei 313,
Sector 6 Bucureşti 060042
Web www.acs.pub.ro

Tel +40 021 402 9179
Fax +40 021 410 1044
Email decanat@acs.pub.ro

UNIVERSITATEA POLITEHNICA BUCURESTI

Facultatea Automatica si Calculatoare

CAIET DE PRACTICĂ

Student: *Dragomir Bianca-Nicola*

Domeniul de studii: *Automatică şi informatică aplicată*

Tema: *Court Judgement Prediction*

Anul de studii: *III* ***Grupa*** *333AC*

Partener de practica: Clementin Cercel

Practica s-a desfasurat in locatia *online*

Stagiul de practica s-a desfasurat in perioada: *27.06.2024 – 09.09.2024*

Numarul total de ore de practica desfasurate: *360*

Activitatea de practica

Saptamana	Activitate	Descrierea activității
S1	Onboarding	În prima săptămână a stagiului, am fost introdus în obiectivele proiectului pe care l-am ales dintr-o listă oferită de profesorul coordonator. Proiectul se concentrează pe aplicarea Inteligenței Artificiale în domeniul juridic, și anume: predicția rezultatelor deciziilor judiciare. Am explorat materialele de referință și platformele relevante care vor fi utilizate pe parcursul stagiului, inclusiv site-uri web ce conțin informații despre hotărâri juridice și articole de specialitate din competiții relevante desfășurate online în ani anteriori (SemEval-2023 Task 6: Court Judgement Prediction with Explanation). În cadrul acestei etape inițiale, mi-am format o bază solidă de cunoștințe care să sprijine activitățile ulterioare.
S2	Cercetare de literatură și familiarizarea cu subiectul	În a doua săptămână, m-am concentrat pe o revizuire extinsă a literaturii de specialitate privind aplicarea Inteligenței Artificiale în domeniul juridic. Am studiat în detaliu conceptele de bază ale Inteligenței Artificiale Explicabile (XAI) și modul în care aceasta poate fi aplicată pentru a oferi transparență și

		interpretabilitate în predicțiile juridice. Am compilat o listă cu lucrări esențiale și studii de caz care explorează utilizarea AI în predicțiile legale și am identificat principalele provocări, cum ar fi disponibilitatea datelor, interpretabilitatea modelului și conformitatea cu reglementările legale.
S3	Cercetare de literatură axată pe modelele de LLMs	Săptămâna a treia a fost dedicată aprofundării studiului modelelor de limbaj mare (LLMs) precum GPT-4, BERT și modele specializate precum LegalBERT. Am analizat în detaliu modul în care aceste modele sunt utilizate pentru sarcini specifice, cum ar fi clasificarea textelor juridice, rezumarea documentelor și predicția rezultatului cazurilor. Evaluarea acestor modele a fost realizată în contextul predicției deciziilor judiciare, punând accent pe factori critici precum acuratețea, scalabilitatea și ușurința integrării cu tehnicile de Inteligență Artificială.
S4	Cercetare de literatură axată și explorarea modelelor LLM existente pentru limba română	În această săptămână, am realizat o revizuire targetată a literaturii pentru a identifica și analiza modelele de limbaj mari existente pentru limba română, cu un accent deosebit pe aplicabilitatea lor la textele juridice. Am evaluat performanța modelelor de limbaj românești, cum ar fi jurBERT (Romanian BERT Model for Legal Judgement Prediction), și am investigat provocările legate de adaptarea acestor modele generale la contextul juridic românesc. Această cercetare a oferit o perspectivă clară asupra modului în care LLM-urile pot fi optimizate pentru a răspunde cerințelor specifice ale limbajului juridic din România.
S5	Configurare pentru scraping (rejust.ro)	În săptămâna a cincea, am început configurarea unui pipeline de scraping pentru platforma rejust.ro, care furnizează decizii judiciare românești. M-am familiarizat cu regulile legate de confidențialitatea datelor și aspectele etice implicate în lucrul cu date juridice. În cadrul acestei activități, am implementat scripturi de scraping folosind Python și biblioteci precum BeautifulSoup, Scrapy și Selenium. Pe măsură ce am avansat, am întâmpinat provocări legate de gestionarea CAPTCHA-urilor și a conținutului dinamic, ceea ce m-a determinat să contactez un alt student recomandat de profesorul

		coordonator, care deține experiență relevantă în acest domeniu, pentru a obține sfaturi și soluții.
S6	Configurare pentru scraping (lege5.ro)	După ce am constatat că platforma rejust.ro nu permite accesul programatic la jurisprudența națională, am decis, la sugestia profesorului coordonator, să utilizez platforma lege5.ro pentru extragerea datelor necesare. Am dezvoltat un pipeline de scraping pentru lege5.ro, concentrându-mă pe extragerea articolelor legale, jurisprudenței și altor documente relevante. În această etapă, am abordat provocări tehnice complexe, cum ar fi paginarea profundă și extragerea de conținut pe mai multe niveluri, asigurând totodată gestionarea eficientă a erorilor și logarea pentru a garanta reproducibilitatea procesului. Am colaborat, de asemenea, cu doi colegi care lucrau pe teme similare pentru a accelera procesul de extragere a datelor.
S7	Conectarea la clusterul UPB, proiectarea setup-ului experimental + scraping	În săptămâna a șaptea, am stabilit conexiunea la clusterul UPB pentru a rula sarcinile de calcul intensiv necesare în procesarea și analiza datelor juridice. Un aspect critic al acestei etape a fost abordarea limitărilor impuse de platforma rejust.ro, care restricționează numărul de accesări la aproximativ 1000 de solicitări per adresă IP. Pentru a depăși această limitare, am implementat un setup experimental care include automatizarea schimbării serverului VPN. Automatizarea schimbării serverului VPN a fost realizată prin scripturi separate, care folosesc biblioteca pyautogui pentru a simula interacțiunile cu interfața ProtonVPN. Prin acest cod, am reușit să mut cursorul mouse-ului la coordonatele specifice ale butonului de schimbare a serverului și să declanșez click-ul automat, astfel asigurând schimbarea adresei IP de fiecare dată când limita de acces a fost atinsă. Pauzele scurte incluse în cod permit serverului să se schimbe complet și să se stabilizeze conexiunea înainte de a relua procesul de scraping. În plus, aceste scripturi au fost integrate în workflow-ul general, permițând continuarea automatizată a colectării datelor chiar și atunci când IP-ul este schimbat. Aceasta

		a permis colectarea unui volum mare de date fără a încălca regulile impuse de platforma rejust.ro.
S8	Rulare webscraping	După ce setup-ul experimental și automatizarea schimbării serverului VPN au fost puse în aplicare, a început procesul de scraping într-un mod automatizat și eficient, asigurând un flux constant de date noi. Loop-ul principal a realizat extragerea datelor juridice în blocuri de câte 100 de pagini. După fiecare bloc procesat, scriptul verifica dacă numărul maxim de solicitări per IP a fost atins. Dacă sesiunea era redirecționată către o pagină de "Access Denied", funcția de schimbare a serverului VPN era apelată automat, după care procesul de scraping era reluat. În plus, funcția task din cod a fost utilizată pentru a accesa fiecare URL specificat și pentru a extrage datele necesare, care au fost apoi curățate și salvate în fișiere CSV. Codul a fost conceput pentru a rezolva erorile comune și pentru a relua automat scraping-ul din punctul în care a fost întrerupt, asigurând astfel o continuitate neîntreruptă a procesului de colectare a datelor.
S9	Compararea performanței modelelor alese și adaptarea modelelor bazate pe drept cu cunoștințe românești	În săptămâna a noua, m-am concentrat pe evaluarea diferitelor tehnici și metode disponibile pentru a îmbunătăți predicția deciziilor judiciare pe baza setului de date juridic românesc. Pe măsură ce am aprofundat în aplicarea modelelor LLM, am realizat că gestionarea și adaptarea LLM-urilor la specificul datelor juridice românești prezintă provocări semnificative, care depășeau resursele și cunoștințele disponibile în acel moment. În urma acestei constatări, am decis să mă reorientez către metode mai accesibile și bine documentate, concentrându-mă pe optimizarea tehnicilor de preprocesare și pe utilizarea modelelor tradiționale de învățare automată. Am continuat procesul de scraping pentru a colecta date relevante, asigurând astfel un set de date robust, pe baza căruia am putut să testez și să validez aceste metode. Am monitorizat cu atenție performanța modelelor utilizate, ajustând parametrii și aplicând tehnici de preprocesare pentru a obține rezultate cât mai precise și relevante.

S10	Concluzie și raport final	În ultima săptămână a stagiului, m-am concentrat pe completarea și finalizarea caietului de practică, conform cerințelor internshipului. Am integrat toate activitățile desfășurate pe parcursul perioadei de practică și m-am asigurat că toate aspectele relevante sunt documentate într-un mod clar și concis.

TUTORE **STUDENT**

DRAGOMIR BIANCA-NICOLA

(Nume, prenume, semnatura) (Nume, prenume, semnătura)