BIOSINF I – IAECV – Proiect

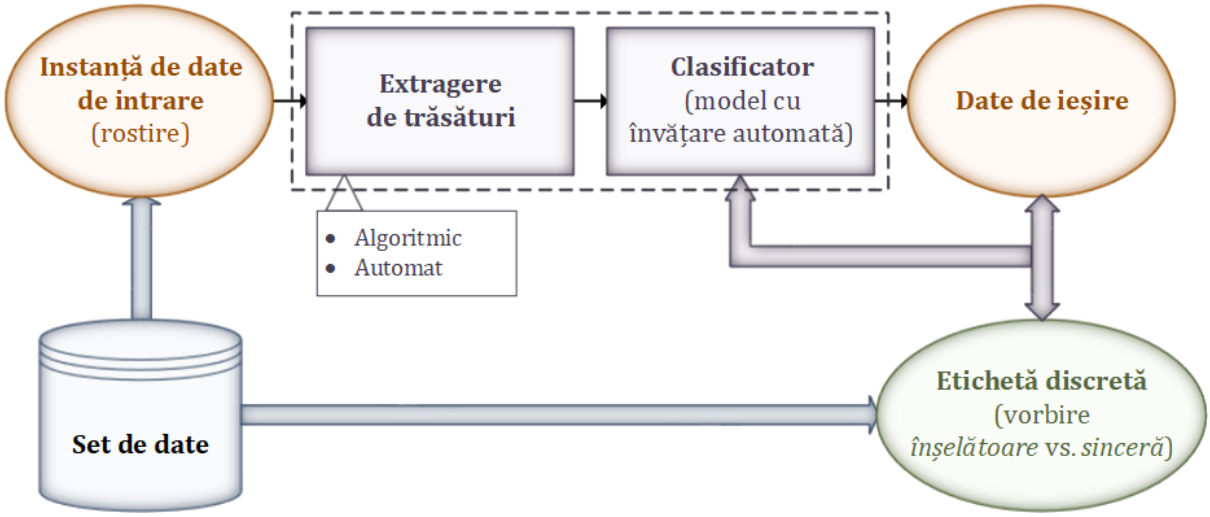
Sistem cu inteligență artificială pentru detecția automată a vorbirii înșelătoare

A. Descrierea proiectului

Obiectivul general al proiectului constă în dezvoltarea unui sistem de învățare automată pentru detecția automată a minciunilor din vorbire (mai precis, a vorbirii *înșelătoare* vs. *sincere*).

Descrierea sistemului

Schema bloc a sistemului este ilustrată în figura de mai jos:



Pentru dezvoltarea sistemului se utilizează setul de date pus la dispoziție în cadrul proiectului, detalierea acestuia fiind făcută în subsecțiunea următoare.

Sistemul cuprinde un bloc de extragere de trăsături și un clasificator.

În cazul extragerii algoritmice a trăsăturilor, cele două blocuri sunt distincte și este vorba de trăsături de semnal vocal precum cele studiate în cadrul disciplinelor **TBASVSB** (*Tehnologii biometrice. Analiza semnalului vocal și a semnalelor biologice*) și **IA1SCIA** (*Inteligență artificială I: sisteme clasice de învățare automată*) – ex: frecvența fundamentală, frecvențele centrale ale formanților, coeficienții Mel-cepstrali, descriptori spectrali, coeficienți Δ și ΔΔ etc. Pentru extragerea acestor trăsături este necesară utilizarea unor pachete Python specializate sau implementarea manuală în Python a unor algoritmi de extragere.

În cazul extragerii automate a trăsăturilor, aceasta este realizată intrinsec de modelul de învățare automată (în particular, rețele neurale convoluționale), deci blocul de extragere de trăsături este parte integrantă a modelului (*simbolizat în figură prin*

chenarul cu linie întreruptă). În această situație trebuie extrase spectrogramele instanțelor audio (rostirile) și oferite modelului ca date de intrare.

Sistemul trebuie să ofere la ieșire predicții asupra apartenenței instanțelor de date la cele două clase luate în considerare (vorbire *înșelătoare* vs. *sinceră*).

Descrierea setului de date

Setul de date cuprinde **121** de înregistrări audio în limba engleză a unor persoane de interes (inculpați, martori etc.) realizate în cadrul unor procese judiciare notorii din Statele Unite ale Americii. Verdictele de vinovăție sau nevinovăție, respectiv de exonerare în unele cazuri a unor persoane anterior private de libertate, au permis adnotarea obiectivă a datelor ca **61** de înregistrări etichetate drept *înșelătoare* și **60** de înregistrări etichetate drept *sincere*. Durata totală a conținutului este de 56 min, cu durata medie a înregistrărilor de 28 s. Excluzând procurorii, avocații și alți intervievatori, numărul total de subiecți (vorbitori) distincți este **56** (**22** de gen feminin, **34** de gen masculin).

Fișierele audio, stocate în format WAV (necomprimat), se găsesc în subdirectorul „extrAudio". Înregistrările au fost deja preprocesate: aduse la frecvență de eșantionare de **16 kHz**, normate în amplitudine și salvate în format PCM pe 16 biți. Cele 61 de înregistrări etichetate ca *înșelătoare* urmează convenția de nume ***trial_lie_xxx.wav***, cu „xxx" între „001" și „061", iar cele 60 de înregistrări etichetate ca *sincere* urmează convenția de nume ***trial_truth_xxx.wav***, cu „xxx" între „001" și „060".

Adnotările înregistrărilor, stocate în format CSV, se găsesc în subdirectorul „datasetAnnotation", numele acestora având o corespondență 1:1 față de fișierele audio. Fiecare fișier CSV conține, **pentru fiecare rostire distinctă prezentă în cadrul înregistrării:** momentul de început, momentul de sfârșit, și identitatea și genul vorbitorului. Momentele de început și de sfârșit ale rostirilor sunt date ca valori numerice (în secunde). Identitățile vorbitorilor sunt codificate prin valori numerice (de la 1 la 56) sau prin valoarea specială „TM" (în cazul procurorilor, avocaților sau al altor persoane care nu reprezintă subiecți în cadrul setului de date). Genurile vorbitorilor sunt codificate prin valorile „F" sau „M".

Numărul total de rostiri incluse în setul de date este **931** (dintre care **464** etichetate drept *înșelătoare* și **467** etichetate drept *sincere*).

Metodologie experimentală

Utilizarea setului de date trebuie să fie la nivel de rostire, deci numărul de instanțe de date (rostirile) este **931**. Se remarcă faptul că cele două clase includ aproximativ același număr de instanțe, deci se pot considera echilibrate.

Pentru extragerea algoritmică a trăsăturilor, fiecare rostire trebuie împărțită în cadre de durată standard (10-30 ms), cu sau fără suprapunere. După ce au fost extrase trăsăturile la nivel de cadru, se aplică funcții statistice pentru a obține valorile acestora

la nivel de instanță (rostire). Pentru extragerea spectrogramelor, trebuie utilizată aceeași granularitate temporală (cadre de aceeași durată) ca în cazul anterior.

Se recomandă descompunerea în cadre de **25 ms** durată (**400** de eșantioane), cu suprapunere de **15 ms** între ele (**240** de eșantioane), deci pas de **10 ms** (**160** de eșantioane), utilizând **ferestre Hamming**.

Pentru extragerea spectrogramelor, în afara parametrilor din paragraful anterior, **se recomandă** calculul DFT în **512** puncte în frecvență și păstrarea conținutului aferent doar intervalului **[0, 8] kHz**, deci aferent doar primelor **257** de puncte în frecvență. De asemenea, **se recomandă** utilizarea scării liniare pentru frecvență și logaritmarea amplitudinilor (logaritm în baza 10, apoi înmulțire cu 10, unitatea de măsură fiind [dB]).

Deoarece toate spectrogramele trebuie să aibă aceeași dimensiune finală, dar rostirile sunt de durate diferite, vectorii de semnal audio aferenți instanțelor de date (rostirilor) trebuie **completați cu eșantioane de 0** înainte de a extrage spectrogramele.

În cazul trăsăturilor extrase algoritmic, trebuie utilizate **cel puțin frecvența fundamentală** și **primii 13 coeficienți Mel-cepstrali (standard)**, dar **se recomandă** mai multe. De asemenea, **se recomandă** utilizarea **mediei** și **deviației standard** ca funcții statistice pentru obținerea valorilor trăsăturilor la nivel de instanță din valorile trăsăturilor la nivel de cadru – Ex: o rostire are durata de 1 s; se descompune în 98 de cadre (cf. recomandării); se extrag cel puțin cele 14 trăsături (cf. recomandării) pentru fiecare dintre cele 98 de cadre; se calculează media și deviația standard pentru cele 98 de valori pentru fiecare dintre cele 14 trăsături, rezultând un vector de 28 de trăsături ce descrie această instanță de date (rostire).

În cazul trăsăturilor extrase algoritmic, la nivel de rostire, valorile acestora trebuie normalizate **pentru fiecare vorbitor distinct** folosind normalizare de tip **z-score**.

Trebuie dezvoltate următoarele tipuri de modele de învățare automată, studiate în cadrul disciplinelor **IA1SCIA** (*Inteligență artificială I: sisteme clasice de învățare automată*) și **IA2RNP** (*Inteligență artificială II: rețele neurale profunde*):

- mașini cu vectori suport (**SVM**);
- ansambluri de arbori de decizie (**RF**);
- rețele neurale complet-conectate (**FCNN**);
- rețele neurale convoluționale (**CNN**).

Datele de intrare pentru modelele de tip **SVM**, **RF**, și **FCNN** sunt trăsăturile extrase algoritmic la nivel de rostire. În cazul rețelelor neurale convoluționale (**CNN**), sunt spectrogramele, extragerea de trăsături făcându-se intrinsec (automat) de către model.

Pentru fiecare tip de model, trebuie realizate experimente folosind diferite configurații obținute prin modificarea **hiperparametrilor**:

- **SVM:** parametrul de regularizare *C*, funcția de transformare (*kernel*) etc.;
- **RF:** numărul de arbori de decizie, adâncimea maximă permisă pentru arbori, funcția criteriu etc.;

- **FCNN:** numărul de straturi ascunse, numărul de neuroni pentru fiecare strat ascuns, funcțiile de activare etc.;
- **CNN:** numărul de straturi convoluționale, numărul de filtre, dimensiunea filtrelor (*kernel*) etc.

Se recomandă dezvoltarea a cel puțin **10** configurații distincte pentru modelele de tip **SVM** și **RF**, a cel puțin **100** de configurații distincte pentru cele de tip **FCNN** și a cel puțin **20** de configurații distincte pentru cele de tip **CNN**.

În cazul modelelor de tip **FCNN** și **CNN**, **se recomandă** ca ultimul strat să conțină **un singur neuron** folosind **funcția sigmoidă** ca funcție de activare. De asemenea, **se recomandă** utilizarea **funcției de entropie încrucișată binară** ca funcție de cost, a algoritmului de optimizare ***Adam***. Este **puternic încurajată** și utilizarea unor tehnici avansate de antrenare, în special antrenarea limitată în timp (*early stopping*), antrenarea selectivă a neuronilor (*dropout*) sau normalizarea activărilor per lot (*batch normalization*), respectiv regularizare (folosind norma L1 sau norma L2).

Se recomandă utilizarea pachetelor Python **scikit-learn** și **Keras/TensorFlow** pentru dezvoltarea modelelor de învățare automată.

Ca metodologie de validare, trebuie folosită **validarea încrucișată** cu **5** împărțiri (5-*fold cross-validation*), folosind **80%** dintre date pentru antrenare și **20%** pentru validare, asigurând o distribuire proporțională atât a instanțelor de date etichetate drept *înșelătoare*, cât și a celor etichetate drept *sincere*. De asemenea, trebuie asigurată **independența relativ la vorbitori** pentru subseturile de antrenare și validare, i.e., toate rostirile aparținând unui vorbitor trebuie să se regăsească exclusiv în subsetul de antrenare sau în subsetul de validare, și trebuie asigurată (pe cât posibil) **împărțirea proporțională a vorbitorilor relativ la gen** pentru subseturile de antrenare și validare, i.e., în ambele subseturi să se regăsească aceeași proporție de vorbitori de gen feminin și de gen masculin.

Pentru evaluarea modelelor, se alege ca măsură de performanță **acuratețea**. Se amintește faptul că, pentru unele tipuri de modele, valorile de ieșire trebuie rotunjite. La final, pe baza tuturor rezultatelor, se determină care este **cel mai performant** model.

B. Organizare și desfășurare

Se lucrează în echipe de câte **3** persoane. Distribuirea sarcinilor în cadrul echipei rămâne la latitudinea membrilor acestora, însă trebuie respectată o încărcare echitabilă a fiecărei persoane (contribuția individuală să fie între **30%** și **40%**).

În prima ședință de proiect sunt prezentate tema de proiect, etapele care trebuie parcurse și modul de evaluare. Următoarele 5 ședințe sunt dedicate discuțiilor legate de dificultăți sau neclarități întâmpinate în etapele de dezvoltare a sistemelor. Proiectul trebuie finalizat și încărcat până la data specificată în secțiunea C, iar evaluarea finală are loc în ultima ședință de proiect.

C. Finalizare și evaluare

- Se creează o arhivă **ZIP** care să conțină:
  - Un document text (TXT) cu lista de pachete Python utilizate.
  - Toate fișierele Python (PY) dezvoltate în cadrul proiectului.
  - O prezentare (PowerPoint/PDF/etc.) a proiectului.
- Arhiva **trebuie** să urmeze convenția de nume:
  IAECV_Echipa_<Nr.> (ex: ***IAECV_Echipa_1.zip***).
- Arhiva se încarcă pe pagina Moodle a disciplinei.
  Termen limită: Duminică, 18.05.2025, 23h59.
- În urma evaluării pachetului de fișiere se acordă nota preliminară (între **1** și **10**), valabilă pentru toți membrii echipei.
- În ultima ședință de proiect (**Miercuri, 28.05.2025**) are loc evaluarea orală a proiectelor, timpul alocat fiind de **10 min**/echipă.
- În urma evaluării orale, fiecare membru al echipei obține un scor (între **0.0** și **1.0**) folosit pentru a scala nota preliminară obținută în etapa anterioară, rezultând nota individuală finală (între **1** și **10**).