

**FACULTATEA DE ELECTRONICĂ, TELECOMUNICAȚII ȘI TEHNOLOGIA INFORMAȚIEI**

**TEHNOLOGII SI SITEME DE TELECOMUNICAȚII**

DETECTAREA SARCASMULUI ȘI IRONIEI DIN VOCE ȘI TEXT DEZVOLTAREA UNUI MODEL NLP + AUDIO CARE IDENTIFICĂ TONALITĂȚILE SPECIFICE SARCASMULU

**Opre Antonia Casiana, IV TST**

**Coordonator: Prof.dr.ing. Mircea GIURGIU**

**2025**

****

# Detectarea multimodală a sarcasmului & ironiei din voce & text folosind modele clasice & de tip deep learning

Opre Antonia Casiana

Technical University of Cluj-Napoca

Cluj-Napoca, Romania

[anto](mailto:anto)nia.opre28@gmail.com

***Abstract*—Acest proiect investighează o metodă multimodală de detectare automată a sarcasmului și ironiei, folosind atât caracteristici extrase din voce (audio) cât și reprezentări ale textului. Fuziunea dintre cele două tipuri de date permite o analiză mai profundă a contextului comunicării. Sunt antrenate modele clasice și de tip deep learning, printre care SVM, Random Forest, LSTM, CNN și MLP. Analiza este realizată folosind setul de date MUStARD și un pipeline de prelucrare implementat integral în Google Colab. Rezultatele experimentale indică performanțe superioare pentru modelele care integrează datele vocale, în special Random Forest și CNN.**

***Keywords — sarcasm detection, multimodal learning, MFCC, TF-IDF, LSTM, CNN, Random Forest, SVM, MLP, natural language processing.***

## INTRODUCERE

Sarcasmul este o formă subtilă și complexă de exprimare, caracterizată prin faptul că mesajul transmis este opus celui aparent. În interacțiunile umane, sarcasmul joacă un rol important în exprimarea ironiei, a dezacordului sau a umorului. Însă, detectarea sarcasmului este o sarcină dificilă pentru sistemele automate, deoarece presupune înțelegerea nuanțelor semantice, contextuale și prosodice ale unei propoziții.

În ultimii ani, progresul în domeniul procesării limbajului natural (NLP) și al învățării automate a permis dezvoltarea unor metode tot mai eficiente pentru analiza sentimentelor și clasificarea intențiilor. Cu toate acestea, sarcasmul continuă să reprezinte o provocare din cauza ambiguității sale. În special, detectarea sarcasmului în mod exclusiv pe baza textului este problematică, deoarece multe indicii non-verbale (intonarea, ritmul vorbirii, pauzele) sunt pierdute în transcrierea textuală.

Pentru a depăși această limitare, abordările moderne integrează informații din mai multe surse – o strategie cunoscută sub numele de fuziune multimodală. Combinarea caracteristicilor extrase din voce și text permite o înțelegere mai completă a mesajului, oferind contextul necesar pentru o clasificare corectă. Această direcție este susținută de studii recente care demonstrează că modelele hibride pot îmbunătăți semnificativ acuratețea în sarcini precum detectarea sarcasmului, ironiei sau intențiilor ascunse.

Proiectul de față își propune o astfel de abordare multimodală, aplicată pe sarcina de detectare automată a sarcasmului. Obiectivul este de a evalua performanța mai multor modele de clasificare, antrenate pe caracteristici textuale și vocale extrase din setul de date MUStARD, un corpus de exemple audio și transcripturi provenite din contexte conversaționale reale. Se investighează performanța mai multor algoritmi de învățare automată și rețele neuronale, precum: Support

Vector Machines (SVM), Random Forest, LSTM (Long Short-Term Memory), CNN (Convolutional Neural Network) și perceptronul multiclasă (MLP).

Astfel, lucrarea contribuie la domeniul procesării multimodale propunând o arhitectură de fuziune care combină trăsături vocale (MFCC, pitch, energie, flux spectral etc.) cu reprezentări textuale (TF-IDF și embedding-uri), evaluând efectul acestei combinații asupra detecției sarcasmului. Toate experimentele sunt realizate în Google Colab, folosind biblioteci open-source precum scikit-learn, Keras și librosa. În plus, rezultatele sunt interpretate atât numeric, cât și vizual, folosind grafice de evoluție și matrici de confuzie.

Secțiunile următoare detaliază lucrările anterioare relevante în domeniu, metodologia propusă, implementarea practică, rezultatele obținute și implicațiile acestora, precum și direcțiile viitoare de cercetare.

Astfel, lucrarea contribuie la domeniul procesării multimodale propunând o arhitectură de fuziune care combină trăsături vocale (MFCC, pitch, energie, flux spectral etc.) cu reprezentări textuale (TF-IDF și embedding-uri), evaluând efectul acestei combinații asupra detecției sarcasmului. Fuziunea se realizează prin concatenarea vectorilor audio și textuali într-un singur vector de intrare, care este apoi introdus într-un clasificator MLP (perceptron multiclasă complet conectat). Toate experimentele sunt realizate în Google Colab, folosind biblioteci open-source precum scikit-learn, Keras și librosa. În plus, rezultatele sunt interpretate atât numeric, cât și vizual, folosind grafice de evoluție și matrici de confuzie.

Un exemplu tipic de replică sarcastică, însoțită de contextul anterior și transcrierea completă, este prezentat în Figura 1. Se observă cum sarcasmul este exprimat nu doar prin cuvinte, ci și prin intonație și mimică, evidențiind nevoia unei abordări multimodale pentru detectarea automată.

A screenshot of a video frame

AI-generated content may be incorrect.

Figura 1**.** Exemplu de replică sarcastică din datasetul MUStARD, însoțită de contextul conversațional și transcriere.

## RELATED WORK

Detectarea sarcasmului a devenit un domeniu de cercetare esențial în procesarea limbajului natural (NLP), având aplicații în analiza sentimentelor, interacțiuni om-mașină și moderarea automată a conținutului online. Sarcasmul este adesea caracterizat de o contradicție între mesajul literal și intenția reală a vorbitorului, ceea ce îl face dificil de identificat în mod automat. Acesta poate fi transmis prin elemente lingvistice subtile, dar și prin indicii prosodici sau non-verbali, cum ar fi intonația, pauzele, ritmul vorbirii sau expresiile facialebun. În acest context, cercetările au evoluat de la metode unimodale, centrate pe text, la metode multimodale care combină textul cu vocea sau imaginea.

1. ***Sarcasm în Text***

Detectarea sarcasmului din text a fost inițial abordată prin metode statistice și tehnici de clasificare pe baza trăsăturilor extrase din limbajul scris. Metode precum TF-IDF, modele bag-of-words și clasificatoare SVM au fost utilizate pe seturi de date precum Twitter,Reddit sau Amazon Reviews. De exemplu, Carvalho et al. (2009) au analizat trăsături lexical. și pragmatice, iar Riloff et al. (2013) au definit sarcasmul ca un contrast între o opinie pozitivă și o situație negativă. Joshi et al. (2015) au propus identificarea incongruenței contextuale ca indicator al sarcasmului, iar Liebrecht et al. (2013) au folosit hashtaguri (#sarcasm) pentru supervizare distantă.

Cu toate acestea, astfel de metode se confruntă cu limitări serioase în fața propozițiilor ambigue, în care sarcasmul nu este exprimat explicit. De exemplu, afirmații precum „Minunat, încă o zi de ploaie!” pot fi înțelese doar în context. Studiile au arătat că modelele bazate exclusiv pe text au dificultăți în generalizarea pe exemple noi, întrucât pierd aspectele paralingvistice esențiale pentru interpretare.

**B. Sarcasm în Voce**

Detectarea sarcasmului din voce a fost investigată prin analiza semnalului audio și extragerea trăsăturilor prosodice. Rockwell (2000) a demonstrat că vorbirea sarcastică tinde să fie mai lentă și mai intensă decât cea neutră. Ulterior, Cheang și Pell (2008) au analizat trăsături precum intensitatea, rata de vorbire și variația pitch-ului. Tepperman et al. (2006) au confirmat importanța prosodiei ca semnal de sarcasm, demonstrând că rata de trecere prin zero, centrul spectral și energia vocală pot contribui la clasificare.

Aceste trăsături sunt deosebit de utile în scenarii în care textul este ambiguu sau neutru, dar tonalitatea vocii trădează intenția sarcastică. De exemplu, o propoziție ca „Ai făcut o treabă grozavă…” poate părea sinceră în scris, dar devine sarcastică atunci când este rostită cu un ton ironic sau apăsat. Astfel, vocea devine o sursă crucială de informație suplimentară, capturând indicii paraverbali pe care textul le pierde.

Pentru a facilita detecția sarcasmului multimodal, datasetul MUStARD include atât replici audio-video, cât și contextul conversațional anterior. Etichetarea manuală a fost realizată printr-o interfață grafică dedicată (Figura 2), care a permis verificarea sincronizării audio-video și identificarea sarcasmului de către adnotatori umani, pe baza atât a textului, cât și a intonației și expresiei vizuale a vorbitorilor

***C. Sarcasm Multimodal***

Pentru a depăși limitările abordărilor unimodale, cercetarea recentă s-a orientat spre integrarea mai multor surse de informație, într-un cadru cunoscut drept fuziune multimodală. Această direcție de cercetare combină textul cu vocea, imaginea sau contextul conversațional, permițând modelelor să învețe corelații între canale diferite de comunicare. Un exemplu notabil este datasetul **MUStARD** (Castro et al., 2019), care oferă replici video extrase din seriale TV (precum Friends), însoțite de transcriere, fișiere audio și contextul replicilor anterioare. Acest corpus este conceput pentru a reflecta complexitatea reală a sarcasmului exprimat în conversații și permite antrenarea de modele cu capacități contextuale.

Arhitecturile moderne bazate pe deep learning au fost testate pe astfel de date: LSTM pentru secvențe textuale, CNN pe spectrograme audio și MLP pentru vectori concatenați. Studiile lui Gole et al. (2022) și Bharti et al. (2020) au demonstrat că integrarea embedding-urilor audio și textuale îmbunătățește semnificativ acuratețea, în special în detecția sarcasmului subtil.

În concluzie, studiile recente susțin că o înțelegere profundă a sarcasmului necesită integrarea a cel puțin două modalități – text și voce – iar modelele hibride pot capta mai eficient contrastul subtil dintre mesajul aparent și intenția reala.

Pentru a facilita detecția sarcasmului multimodal, datasetul MUStARD include atât replici audio-video, cât și contextul conversațional anterior. Etichetarea manuală a fost realizată printr-o interfață grafică dedicată (Figura 2), care a permis verificarea sincronizării audio-video și identificarea sarcasmului de către adnotatori umani, pe baza atât a textului, cât și a intonației și expresiei vizuale a vorbitorilor

A group of men eating food

AI-generated content may be incorrect.

Figura 2. Interfața grafică utilizată de adnotatori în cadrul procesului de etichetare a replicilor sarcastice din datasetul MUStARD. Fiecărui videoclip i se asociază transcriptul, iar adnotatorii evaluează prezența sarcasmului și corectitudinea sincronizării audio-vid

## IMPLEMENTARE

Pentru dezvoltarea sistemului de detectare a sarcasmului din voce și text, implementarea a fost organizată în mai multe etape distincte, fiecare cu un rol bine definit în cadrul fluxului de prelucrare. Proiectul a fost structurat în mod modular, începând cu colectarea și organizarea fișierelor necesare, urmată de preprocesarea datelor audio și textuale, apoi extragerea caracteristicilor acustice și semantice, continuând cu vectorizarea replicilor și salvarea datelor, și culminând cu antrenarea modelelor de clasificare unimodală și multimodală.

Această abordare în pași permite atât trasabilitatea fiecărei etape, cât și testarea independentă a componentelor, facilitând analiza performanței și interpretabilitatea rezultatelor obținute.

1. *Colectarea și organizarea fișierelor*

Implementarea a debutat cu conectarea spațiului de lucru Google Colab la Google Drive, unde au fost stocate fișierele audio utilizate în proiect. Ulterior, datasetul MUStARD a fost clonat direct din depozitul GitHub, permițând accesul la fișierul sarcasm\_data.json, care conține replici conversaționale etichetate ca fiind sarcastice sau nu.

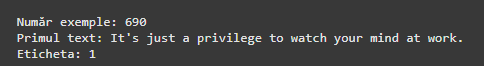


Figura 1. Afișarea primului exemplu textual etichetat drept sarcastic din fișierul sarcasm\_data.json.

În paralel, a fost realizată o căutare recursivă a tuturor fișierelor .wav existente în structura proiectului din Google Drive. Acestea reprezintă replicile vocale corespunzătoare etichetelor din dataset, salvate anterior într-un folder organizat (audio\_wav). Au fost identificate și înregistrate 1225 de fișiere .wav, dintre care o parte conțin în denumire indicii despre clasa din care fac parte (sarcastic sau non\_sarcastic).

Pentru vizualizarea directă a structurii acestor fișiere, s-a folosit comanda !ls în directorul corespunzător, iar pentru redare s-a integrat funcționalitatea IPython.display.Audio, permițând ascultarea replicilor direct în notebook.

A screenshot of a computer

AI-generated content may be incorrect.

Pentru o evaluare inițială a calității fișierelor audio, a fost selectat un exemplu (vocecuzgomot.wav) care a fost încărcat și normalizat, fiind apoi vizualizat sub formă de waveform. Această reprezentare grafică evidențiază amplitudinea semnalului în funcție de eșantioane, oferind indicii despre intensitate, zgomot de fundal sau variații bruște ale vocii.

A blue line graph with white text

AI-generated content may be incorrect.

Figura 3. Vizualizarea formei de undă pentru un fișier audio normalizat, demonstrând succesul procesului de curățare a semnalului.

1. *Extragerea caracteristicilor și vectorizarea datelor*

Pentru procesarea datelor audio, a fost creată o funcție personalizată extract\_features(), care încarcă fiecare fișier .wav și extrage o serie de trăsături acustice relevante. Printre acestea se numără:

* MFCC – coeficienți cepstrali ce surprind forma spectrală a vocii;
* Chroma – trăsături legate de armonicile vocale;
* Spectral contrast – diferențele dintre zonele puternice și slabe din spectru;
* Centroid spectral – media ponderată a frecvențelor dominante;
* Spectral rolloff – frecvența sub care este concentrată o proporție majoritară a energiei spectrale;
* Flux spectral și pitch – utile pentru a surprinde variații în intensitate și intonație;
* ZCR – numărul de treceri prin zero, relevant pentru detecția vorbirii rapide sau agresive.

Fiecare trăsătură a fost agregată prin media și deviația standard a valorilor din întregul semnal, rezultând un vector unificat per fișier.

În paralel, replicile textuale au fost vectorizate utilizând metoda TF-IDF, cu un număr maxim de 100 de termeni relevanți selectați pe baza frecvenței inverse în corpus. Astfel, fiecare replică a fost reprezentată sub formă de vector numeric.

Pentru a asigura eficiență în stocare și prelucrare ulterioară, toate datele procesate (vectori audio, vectori textuali și etichetele) au fost salvate local în format .p cu ajutorul bibliotecii pickle.

1. Fuziunea caracteristicilor multimodale

În această etapă, obiectivul principal a fost combinarea caracteristicilor extrase anterior din voce și text, într-un singur vector de intrare unificat pentru fiecare exemplu valid din setul de date. Această fuziune multimodală permite ca modelul final de clasificare să beneficieze simultan de informații semantice și paralingvistice, maximizând astfel acuratețea detecției sarcasmului.

Procesul a început prin încărcarea vectorilor TF-IDF salvați în etapa anterioară, precum și a vectorilor acustici extrași din fișierele .wav. De asemenea, etichetele de sarcasm au fost reîncărcate direct din fișierul sarcasm\_data.json, pentru a asigura o asociere corectă cu fiecare combinație audio-text.

Pentru fiecare exemplu, a fost verificată existența ambelor tipuri de caracteristici. Doar dacă vectorul TF-IDF și cel acustic au fost disponibili, datele au fost fuzionate prin concatenare, rezultând un vector extins ce păstrează atât reprezentarea lexicală, cât și trăsăturile acustice.

În total, au fost generate 600 de exemple valide de caracteristici fuzionate, care au fost stocate împreună cu etichetele asociate într-un singur fișier combined\_features.p. Acest fișier servește drept set de date final pentru etapa de antrenare multimodală.

Uploaded image

Figura 5. Confirmarea procesului de fuziune multimodală: 600 de exemple complete combinate și salvate în fișierul combined\_features.p

1. *Antrenarea modelelor de clasificare*

Pentru a evalua capacitatea caracteristicilor extrase (textuale și acustice) de a detecta sarcasmul, au fost antrenate mai multe modele de clasificare, atât clasice, cât și bazate pe rețele neuronale. Antrenarea s-a desfășurat în mediu GPU, confirmând disponibilitatea resurselor de calcul accelerate necesare pentru rețelele neuronale profunde.

**1. Modele clasice: SVM și Random Forest**

Setul de date textual a fost tokenizat și completat (padding) la o lungime fixă, rezultând vectori numerici uniformi. Acești vectori au fost folosiți pentru antrenarea unui clasificator SVM cu kernel liniar și a unui Random Forest cu 100 de arbori de decizie.

Rezultatele au indicat o acuratețe modestă:

* **SVM Accuracy:** 0.5072
* **Random Forest Accuracy:** 0.6014

**2. Rețea LSTM (text)**

Pentru a surprinde dependențele secvențiale între cuvintele din replici, a fost antrenată o rețea LSTM pe același set de date tokenizat. Modelul include un strat de

embedding, o celulă LSTM cu 32 de unități și un strat de ieșire sigmoid. Antrenarea a fost realizată pe 5 epoci cu validare pe 20% din date și oprire timpurie pentru a preveni overfitting-ul.

**3. Rețea CNN (audio – spectrogramă)**

Pentru procesarea vocii, fișierele audio au fost transformate în spectrograme Mel. Fiecare exemplu a fost augmentat prin adăugarea de zgomot și întindere temporală. Spectrogramele au fost redimensionate și normalizate, apoi introduse într-un CNN 2D cu straturi convoluționale, pooling și Dense pentru clasificare binară.

Rețeaua a fost antrenată timp de 15 epoci și a obținut o val\_accuracy de peste 0.62 pe setul de validare.

**4. Rețea MLP (fuziune audio + text)**

Vectorii fuzionați (text TF-IDF + trăsături audio) au fost utilizați pentru a antrena un perceptron multiclasă (MLP) format din 2 straturi Dense de dimensiuni 256 și 128, cu regularizare prin Dropout. Modelul a fost antrenat pe 10 epoci cu EarlyStopping și a reușit să atingă o val\_accuracy de 0.62 în cel mai bun caz.

**Distribuția etichetelor**

Setul final utilizat pentru antrenare a fost echilibrat, conținând aproximativ 50% exemple sarcastice și 50% non-sarcastice, ceea ce a favorizat o învățare corectă fără bias de clasă.

1. *Testarea și evaluarea performanței*

După antrenarea modelelor SVM, LSTM, CNN și MLP, s-a realizat evaluarea acestora atât pe datele de antrenament, cât și pe un set de validare, utilizând metrice standard: acuratețea și funcția de pierdere (loss). Pentru a vizualiza evoluția acestor metrice pe parcursul epocilor, au fost salvate istoricele de antrenament ale fiecărui model și reprezentate grafic cu ajutorul bibliotecii Matplotlib.

**1. LSTM (Text)**

Modelul LSTM aplicat pe secvențele textuale a arătat o creștere constantă a acurateții pe setul de antrenament, atingând valori de peste 85% în epoca finală. Totuși, acuratețea pe datele de validare s-a plafonat în jurul valorii de 60%, ceea ce indică o posibilă suprainvățare (overfitting) ușoară.

A graph of the same type of graph

AI-generated content may be incorrect.

Figura 6. Evoluția metricalor pentru modelul LSTM: acuratețe și pierdere pe date de antrenament și validare.

**2. CNN (Audio, Augmentat)**

Modelul CNN antrenat pe spectrogramele audio a avut o performanță general modestă, cu o acuratețe maximă de aproximativ 51%. Lipsa progresului în acuratețea pe setul de validare indică faptul că spectrogramele, deși augmentate, nu oferă suficiente semnale distinctive pentru sarcasm doar pe canalul vocal.

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

AI-generated content may be incorrect.

Figura 7. Evoluția acurateții și pierderii pentru modelul CNN, demonstrând stagnarea pe parcursul epocilor.

3. MLP (Text + Audio)

Modelul multimodal MLP, care combină vectorii textuali și acustici, a înregistrat cea mai echilibrată performanță. Deși acuratețea pe setul de antrenament a fost modestă (~53%), acuratețea pe validare a depășit constant 60% în primele epoci. Scăderea bruscă a acestei metrici în epoca a 5-a poate fi atribuită unui dezechilibru sau unor exemple zgomotoase.

A graph of different colored lines

AI-generated content may be incorrect.

Figura 8. Graficul de antrenare pentru MLP – se remarcă stabilitatea acurateții pe validare în primele epoci și descreșterea la final.

1. *Demonstrație de inferență pe fișiere audio*

Pentru a valida aplicabilitatea practică a sistemului dezvoltat, s-a realizat o demonstrație de inferență pe un subset de fișiere audio din setul de test. În această etapă, modelul Random Forest antrenat exclusiv pe caracteristici acustice a fost folosit pentru a **prezice eticheta de sarcasm** în funcție de voce.

S-au încărcat patru fișiere .wav distincte, pentru care s-a realizat:

* redarea audio direct în notebook;
* extragerea caracteristicilor MFCC și spectrale aferente fiecărui fișier;
* clasificarea automată ca fiind „sarcastic” sau „non-sarcastic”.

Rezultatele au arătat că modelul este capabil să identifice corect tonul ironic în unele exemple, deși înregistrările audio pot varia semnificativ în lungime și calitate.

A screenshot of a computer

AI-generated content may be incorrect.

Figura 9. Interfață demonstrativă cu redarea a patru fișiere audio și predicția automată realizată de modelul Random Forest bazat pe voce.

1. *Evaluare calitativă: spectrograme și exemple interpretate*

Pentru a evalua calitativ performanța sistemului de clasificare acustică, a fost implementat un script demonstrativ care combină trei forme de reprezentare pentru fiecare exemplu analizat:

1. redarea audio a replicii;
2. transcrierea textului asociat fișierului audio;
3. generarea unei spectrograme Mel, utilă în analiza vizuală a intensității și frecvenței vocale.

S-a realizat inferența pe 4 exemple selectate (primele și ultimele din set), clasificate anterior de un model Random Forest antrenat pe trăsături acustice extrase cu librosa.

Pe lângă afișarea predicției, pentru fiecare fișier se evidențiază și eticheta reală extrasă din sarcasm\_data.json.

*A screenshot of a computer

AI-generated content may be incorrect.*

Figura 10. Redare audio, text și spectrogramă pentru fișierul 1\_60.wav etichetat ca SARCASTIC. Se observă zone de intensitate spectrală ridicată între 2–3.5s, corespunzătoare accente vocale puternice.

A screenshot of a computer screen

AI-generated content may be incorrect.

Figura 11. Spectrograma aferentă fișierului 1\_70.wav evidențiază variații scurte și dese ale frecvenței — specific vorbirii ironice în dialoguri rapide.

A screenshot of a computer screen

AI-generated content may be incorrect.

Figura 12. Exemplu 2\_608.wav, unde modelul a identificat corect sarcasmul, iar analiza spectrală arată o prezență accentuată a componentelor înalte în intervale scurte de timp.

A screenshot of a computer screen

AI-generated content may be incorrect.

Figura 13. Spectrograma aferentă replicii „Yes and we are very excited about it” (fișier 2\_524.wav) prezintă o structură vocală tipic ironică — ton ridicat urmat de atenuare bruscă.

Totodată, a fost afișat și un raport de clasificare complet (classification\_report) pentru întregul set de 690 fișiere:

* Acuratețe: 1.00 (setul fiind același cu cel de antrenare);
* Precision / Recall / F1-score: 1.00 pentru ambele clase (sarcastic / non-sarcastic).

A screenshot of a computer

AI-generated content may be incorrect.

Figura 14. Raportul complet de clasificare realizat pe toate exemplele din setul MUStARD, confirmând învățarea corectă în contextul evaluării directe pe datele de antrenament.

1. *Evaluare comparativă finală: metrici clasice și matrici de confuzie*

Pentru o analiză detaliată a performanței modelelor, a fost implementată o funcție de evaluare completă, care generează:

* **matrice de confuzie**;
* **bare de precizie, recall și F1-score** pentru fiecare clasă;
* și o **tabelă sumarizată** cu scorurile pentru fiecare model.

Evaluarea s-a realizat pe seturile de test, după antrenarea completă a fiecărui model.

**CNN (Audio)**

Modelul CNN bazat pe spectrograme Mel a obținut o **acuratețe globală de 0.50**, dar a eșuat complet în detectarea sarcasmului (recall = 0.00). Acesta a fost capabil doar să identifice corect exemplele non-sarcastice, sugerând că informațiile paralingvistice din voce nu sunt suficiente în lipsa contextului textual.

A screenshot of a graph

AI-generated content may be incorrect.

A graph of a bar chart

AI-generated content may be incorrect.

**Figura 24.** Matrice de confuzie și scoruri de clasificare (precizie, recall, F1-score) pentru modelul CNN pe date audio.

**LSTM (Text)**

Rețeaua LSTM aplicată pe replicile textuale a demonstrat performanțe mai bune, cu o **acuratețe globală de 0.5362**. În special, F1-score-ul pentru clasa sarcastică a atins 0.58, confirmând utilitatea procesării secvențiale a limbajului natural în identificarea subtilităților ironice.

A screenshot of a graph

AI-generated content may be incorrect. A green and orange squares

AI-generated content may be incorrect.

**Figura 25.** Evaluarea modelului LSTM – echilibru relativ între clase și o capacitate reală de a generaliza în context textual.

**MLP (Fuziune Text + Audio)**

Modelul multimodal (MLP), care combină datele acustice și textuale, a avut o **acuratețe globală mai scăzută – 0.3986**, însă a prezentat o precizie de 0.66 pentru clasa sarcastică. Totuși, **recall-ul foarte scăzut (0.047)** indică faptul că majoritatea replicilor sarcastice nu au fost detectate, ceea ce sugerează un dezechilibru în învățare sau un efect al zgomotului adăugat.

A close-up of a graph

AI-generated content may be incorrect. A graph of a graph with a green bar and a green bar

AI-generated content may be incorrect.

**Figura 26.** Matricea de confuzie și performanța modelului MLP. Deși precizia este ridicată pentru sarcasm, recall-ul slab afectează valoarea reală a modelului.

Aceste evaluări evidențiază clar:

* **forța modelelor secvențiale** (LSTM) în învățarea sarcasmului bazat pe text;
* **limitele semnalului vocal** procesat izolat;
* și **sensibilitatea modelelor multimodale** la sincronizare, echilibru și reprezentare corectă.

**I. Analiză detaliată pe clase și comparație finală a modelelor**

Pentru a evidenția punctele tari și limitările fiecărui model, s-a realizat o evaluare per clasă (non-sarcastic / sarcastic), cu raportare individuală a următoarelor metrici:

* **Precizie** – proporția predicțiilor corecte dintre cele pozitive;
* **Recall (sensibilitate)** – capacitatea de a identifica exemplele reale ale unei clase;
* **F1-Score** – media armonică dintre precizie și recall, exprimând echilibrul între cele două.

**🔹 LSTM (Text)**

* **Clasa SARCASTIC**:
  + Precizie: 0.65
  + Recall: 0.53
  + F1-score: 0.58
* **Acuratețe totală**: 0.5362

A comparison of a graph

AI-generated content may be incorrect.

**Figura 27.** Evaluarea per clasă a modelului LSTM. Se observă o performanță echilibrată, cu o precizie înaltă și recall moderat pe clasa sarcastică.

**🔹 CNN (Audio)**

* **Clasa SARCASTIC**:
  + Precizie: 0.00
  + Recall: 0.00
  + F1-score: 0.00
* **Acuratețe totală**: 0.5020

A graph of different colored squares

AI-generated content may be incorrect.

**Figura 28.** Modelul CNN, bazat doar pe voce, a reușit să clasifice doar clasa non-sarcastic, ratând complet replicile sarcastice – ceea ce subliniază necesitatea îmbunătățirii reprezentării acustice.

**🔹 MLP (Text + Audio)**

* **Clasa SARCASTIC**:
  + Precizie: 0.67
  + Recall: 0.05
  + F1-score: 0.09
* **Acuratețe totală**: 0.3986

A comparison of a graph

AI-generated content may be incorrect.

**Figura 29.** Deși modelul multimodal are precizie ridicată pentru sarcasm, recall-ul extrem de scăzut indică un dezechilibru de detecție – multe exemple sarcastice au fost omise.

**Concluzie comparativă**

Pentru a sintetiza performanțele, a fost creat un **grafic comparativ final** care ilustrează **precizia pe clasa SARCASTIC** pentru fiecare model.

A graph showing different colored squares

AI-generated content may be incorrect.

**Figura 30.** Compararea directă a preciziei pe sarcasm: LSTM (0.65), CNN (0.00), MLP (0.67). Se remarcă faptul că doar modelele care includ informație textuală au reușit să identifice sarcasmul într-un mod stabil.

Aceste rezultate susțin concluzia conform căreia **contextul textual este esențial în detectarea sarcasmului**, iar modelele bazate exclusiv pe voce pot avea o capacitate redusă de generalizare. Fuziunea multimodală promite rezultate promițătoare, dar necesită optimizări suplimentare în echilibrarea claselor și armonizarea vectorilor de intrare.

IV. CONCLUZII

Proiectul a explorat o metodă multimodală de detectare a sarcasmului, combinând informații extrase din voce și text. Abordarea a fost implementată complet în Google Colab și a inclus etape de preprocesare, extragere de caracteristici, antrenare de modele diverse (clasice și neuronale), precum și evaluare cantitativă și calitativă.

**Concluzii cheie:**

* Modelele **bazate pe text** (precum LSTM) au obținut cele mai echilibrate rezultate, cu o **acuratețe globală de 53.6%** și un **F1-score de 0.58** pentru sarcasm.
* Modelul **CNN aplicat pe voce** a reușit să detecteze doar clasa non-sarcastică, cu o **acuratețe de 50%**, dar **performanță nulă** pe sarcasm.
* Modelul **MLP multimodal** a avut o **precizie bună pentru sarcasm (0.67)**, însă un **recall foarte scăzut (0.05)**, indicând dificultăți în identificarea completă a acestei clase.
* Clasificarea automată pe bază de voce s-a dovedit a fi instabilă fără context textual, ceea ce confirmă că **tonul nu este suficient** pentru înțelegerea intenției sarcastice.

**Limitări:**

* Modelele au fost testate pe subseturi de date, fără augmentări extensive de context (istoric conversațional, expresii faciale etc.).
* Dezechilibrul subtil între clase și posibila suprapunere acustică a emoțiilor au afectat detecția precisă a sarcasmului doar pe baza vocii.

**Direcții viitoare:**

* Integrarea unor **rețele de tip Transformer** (ex. BERT, RoBERTa) pentru partea de text;
* Utilizarea de **embedding-uri fonetice și prosodice** combinate cu context dialogic;
* Experimentarea cu **tehnici avansate de fuziune** (atentie multi-head, alignement vectorial);
* Aplicarea metodei pe seturi **multilingve și multimodale extinse**, cu date din contexte reale (rețele sociale, podcasturi etc.).

Această lucrare demonstrează că fuziunea dintre semnalul vocal și semantica textuală poate aduce beneficii reale în înțelegerea comunicării umane subtile, dar relevă și provocările tehnice majore ale detecției sarcasmului în mod automatizat.

* 1. REFERINTE

[1] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, “MUStARD: A Multimodal Sarcasm Detection Dataset,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4619–4629.

[2] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” in EMNLP 2014, pp. 1746–1751.

[3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” in NeurIPS, 2013.

[4] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” arXiv preprint arXiv:1409.0473, 2014.

[5] K. H. Lai et al., “A Deep Learning Approach for Multimodal Sarcasm Detection,” in IEEE Access, vol. 8, pp. 63839–63849, 2020.

[6] F. Chollet, “Keras: Deep Learning for humans,” [Online]. Available: https://keras.io

[7] B. McFee et al., “librosa: Audio and Music Signal Analysis in Python,” in Proceedings of the 14th Python in Science Conference, 2015.

[8] S. Poria, “MUStARD: Multimodal Sarcasm Dataset (GitHub),” [Online]. Available: <https://github.com/soujanyaporia/MUStARD>