Universitatea Ștefan cel Mare Suceava

Facultatea de Inginerie Electrică și Știința Calculatoarelor

REZUMAT

CERCETĂRI PRIVIND METODE, TEHNICI ȘI APLICAȚII DE TEXT MINING

Domeniul Calculatoare și Tehnologia Informației

Conducător științific:

***Conf. univ. dr. ing.* Mirela DANUBIANU**

Doctorand:

***Ing.* Mihaela MILEA (CHISTOL)**

Suceava, România
2024

UNIUNEA EUROPEANĂ
Proiect cofinanțat din Fondul Social European prin Programul Operațional Capital Uman 2014-2020

Această lucrare a beneficiat de suport financiar prin proiectul

Excelența academică și valori antreprenoriale - sistem de burse pentru asigurarea oportunităților de formare și dezvoltare a competențelor antreprenoriale ale doctoranzilor și postdoctoranzilor – ANTREPRENORDOC

Contract nr: 36355/23.05.2019 POCU/380/6/13
Cod SMIS: 123847
Axa prioritară 6 **-** Educație și competențe
Obiectiv specific 6.13 **-** Creșterea numărului absolvenților de învățământ terțiar universitar și nonuniversitar care își găsesc un loc de muncă urmare a accesului la activități de învățare la un potențial loc de muncă/cercetare/inovare, cu accent pe sectoarele economice cu potențial competitiv, identificate conform SNC, și domeniile de specializare inteligentă, conform SNCDI.

Beneficiar: Parteneri:
Universitatea Ștefan cel Mare Suceava

CUPRINS

Glosar **6**

1 Introducere **7**

1.1 Structura tezei de doctorat 9
1.2 Contribuțiile tezei de doctorat 10
1.3 Articole științifice publicate 12

2 Stadiul actual al cercetărilor în domeniul text mining **13**

2.1 Conceptul text mining 13
2.2 Metode și tehnici de text mining 13
2.2.1 Metode și tehnici de preprocesare date text 14
2.2.2 Metode și tehnici de extragere caracteristici 15
2.3 Principii și tehnici de învățare automată în text mining 16
2.3.1 Abordări fundamentale de învățare automată 16
2.3.2 Tehnici de învățare automată în text mining 16

3 Text mining în domeniul sănătății și asistenței medicale **18**

3.1 Context 18
3.2 Tehnologii utilizate în diagnosticarea tulburării de spectru autist 18
3.3 Text mining în tulburarea de spectru autist 20

4 Percepția utilizatorilor față de diagnosticarea tulburării de spectru autist asistată de tehnologie **22**

4.1 Context 22
4.2 Experiment 22
4.3 Rezultate 24

5 Text mining pentru diagnosticarea tulburării de spectru autist **28**

5.1 Context 28
5.2 Metodologie de lucru 28
5.3 Experiment 30
5.4 Rezultate 35

6 Autism AI Advisor for Diagnosis: sistem de suport pentru diagnosticarea precoce a tulburării de spectru autist **37**

6.1 Context 37
6.2 Metodologie de lucru 37
6.3 Experiment 45
6.4 Rezultate 48

7 Concluzii **52**

Proiecte de cercetare **54**

Referințe bibliografice **55**

GLOSAR

Termen	Definiție
Clasă	Categorie specifică pe care un algoritm o identifică și o atribuie unei anumite instanțe.
Clasificare	Tehnică folosită pentru a prezice informații similare pe baza valorilor claselor preexistente. [1]
Clustering	Tehnică folosită pentru a grupa instanțele similare în grupuri. Tehnica împarte setul de date în subseturi, astfel încât datele din fiecare subset sunt mai apropiate conform unei anumite măsuri de distanță [2].
Corpus	Colecție de documente.
Data mining	Procesul de extragere a cunoștințelor din date organizate în baze de date.
Etichetă	Categorii sau clase asociate unor instanțe de date.
Framework	Ansamblu standardizat de metode și practici care oferă o structură predefinită pentru rezolvarea problemelor cu conotații similare.
Normalizare	Tehnica de a reprezenta valorile în același domeniu.
Tehnologia informației	Tehnologia necesară pentru prelucrarea informației, în particular prin folosirea computerelor.
Text mining	Procesul de extragere a cunoștințelor din date de tip text.
Tulburarea de spectru autist	Deficiența de dezvoltare a creierului, cu implicații în ceea ce privește comunicarea, comportamentul și interacțiunea socială.
Zettabyte	Unitate digitală de măsură. Un zettabyte este egal cu un sextilion de octeți sau 1021 de octeți [3].
Zgomot	O instanță de date care reprezintă o valoare anormală, care deviază semnificativ de la distribuția generală a datelor.

1 INTRODUCERE

În epoca informațională, oamenii folosesc tehnologii pentru a-și amplifica capacitățile fizice, pentru a îndeplini sarcini și pentru a comunica [4]. Comunicarea este parte integrată a vieții de zi cu zi, care indiferent de modul în care se realizează, fizic sau virtual, generează date. Peste 8 miliarde de oameni [5] contribuie la cantitatea uriașă de date consumată la nivel global. Progresele înregistrate în industria tehnologiei informației (IT) oferă mijloace eficiente de creare, colectare și stocare a acestor date. Tendința observată în Figura 1.1 este exponențială și sugerează că în 2025 volumul datelor va ajunge la peste 180 zettabyte [6]. În prezent, datele înseamnă mult mai mult decât cifre și litere, acestea includ imagini, sunete și text. Conform rezultatelor cercetării lui T. King [7] 80% din datele de pe glob sunt în format nestructurat. Această statistică subliniază importanța tehnicilor de procesare a datelor nestructurate precum este text mining.

Text mining este un concept nou al cercetării informatice care contribuie la rezolvarea crizei informaționale prin extragerea automată a cunoștințelor din cantități mari de date nestructurate. Text mining combină metode și tehnici din data mining (DM), învățarea automată (*en.*: machine learning (ML)) și procesarea limbajului natural (*en.*: natural language processing (NLP)). Spre deosebire de data mining, text mining nu a fost „născut" digital, ci a evoluat de la originile sale în revizuirea manuală a manuscriselor antice [8] pentru a deveni un domeniu sofisticat, care utilizează metode de inteligență artificială (*en.*: artificial intelligence (AI)) pentru analiza datelor text. E.L. Tonkin [8] prezintă povestea unui preot iezuit pe nume Roberto Busa, cunoscut astăzi ca fondatorul informaticei umaniste, care a explorat opera Sfântului Toma de Aquino cu scopul de a crea un index al lucrării lui Aquino. În 1949, Busa a călătorit în Statele Unite ale Americii căutând o alternativă automatizată pentru a înlocui cele 10000 de fișe scrise de mână generate deja de lucrarea sa [8]. Această sarcină era imposibilă pentru orice tehnologie a acelor vremuri, motiv pentru care primise refuzuri din partea a 25 de universități americane. Încurajat de sloganul companiei IBM „*The difficult we do right away; the impossible takes a little longer.*", Busa reușește să obțină suportul fondatorului IBM, Thomas J. Watson, și astfel să fie dezvoltat primul proces de explorare automată a textului [8]. Explorarea automată a textului s-a dovedit a fi prea utilă pentru ca potențialul său de afaceri să fie ignorat [8]. Multe companii și medii academice au cercetat această tehnologie de-a lungul timpului. În anii 1990, cercetătorii au început să aplice tehnici statistice și de învățare automată pentru a analiza textul. Odată cu creșterea big data și cloud computing în anii 2000, text mining a evoluat într-un domeniu sofisticat utilizând învățarea profundă (*en.*: deep learning (DL)) pentru a analiza seturi de date vaste în timp real. În prezent, text mining contribuie la realizarea a ceea ce părea impo-

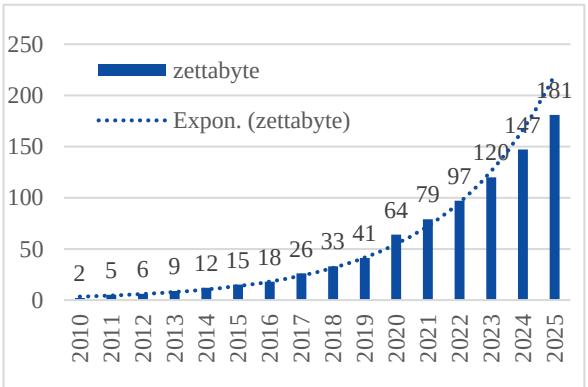


Figura 1.1 Volumul de date la nivel mondial exprimat în zettabyte pe an [6].

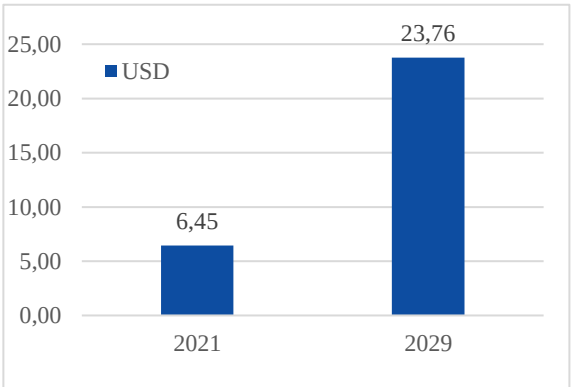


Figura 1.2 Dimensiunea pieței de text mining exprimată în miliarde de dolari pe an [9].

sibil, și anume la atribuirea computerului capacitatea de înțelege limbajul uman. În 2014, programul „Eugene Goostman" a devenit primul software autonom care a reușit să treacă testul matematicianului Alan Turing [10].

Piața de text mining înregistrează o creștere semnificativă datorită necesității de a obține informații din date text nestructurate și interesului crescut pentru analiza volumelor mari de date. Text mining permite companiilor să obțină informații valoroase din feedbackul clienților, să identifice noi oportunități de produse și să monitorizeze percepția brandului pe platformele de social media. Creșterea pieței, este de asemenea determinată de nevoia de informații precise, care să poată fi obținute în timp real, iar companiile investesc tot mai multe resurse în cercetarea și dezvoltarea sistemelor de luare de decizii. Industrii precum cea medicală, comerțul, telecomunicațiile și finanțele folosesc text mining pentru a înțelege satisfacția clienților, pentru a îmbunătăți operațiunile, prezice tendințele emergente și pentru a lua decizii care vor genera creșterea veniturilor. Conform cercetărilor de piață realizate de Data Bridge Market Research (DBMR), se estimează că piața de text mining va evolua la o rată de creștere anuală compusă (CAGR) de 17.70% în perioada de prognoză, urmând să atingă 23.76 miliarde de dolari până în 2029 (Figura 1.2) [9].

Într-o epocă în care informațiile sunt generate într-un ritm fără precedent, capacitatea de a extrage cunoștințe din datele text nu a fost niciodată mai importantă. Relevanța text mining se datorează în special digitalizării constante a tuturor aspectelor vieții umane care este susținută prin reglementări legislative. În 2021, Comisia Europeană a comunicat planul pentru realizarea transformării digitale a economiei și societății europene (SE) numit „*Busola Digitală 2030: calea europeană pentru Deceniul Digital*" [11]. Busola indică patru puncte cardinale pentru această traiectorie: competențe digitale, infrastructuri digitale sigure și durabile, transformarea digitală a întreprinderilor și digitalizarea serviciilor publice [12]. Busola digitală urmărește să creeze un ecosistem sigur centrat pe om, ceea ce va implica modificări majore în domenii cheie. Printre domeniile care vor suferi cele mai mari schimbări este domeniul medical, deoarece se urmărește digitalizarea completă a serviciilor medicale în toate țările membre ale Uniunii Europene (EU) prin furnizarea de servicii de e-sănătate (*en.*: e-health) și acces la dosare electronice de sănătate (*en.*: e-records) [13]. Comisia europeană, estimează că trecerea de la servicii pe hârtie la servicii online va genera beneficii de până la 120 de miliarde de euro pe an în întreaga Europă [13]. Cu toate acestea, obiectivele ambițioase vor fi greu de atins dacă nu se depun eforturi și pentru dezvoltarea produselor software capabile să proceseze volumele mari de date rezultate în urma digitalizării.

Organizația Mondială a Sănătății (OMS) estimează că 1 din 10 oameni suferă din cauza erorilor medicale, iar anual peste 3 milioane de decese sunt cauzate de îngrijiri nesigure [14]. Este pe deplin la îndemâna inginerilor să creeze tehnologii care procesând automat datele pacienților să reducă volumul erorilor medicale. În timp ce tehnologiile nu pot înlocui niciodată medicii, acestea pot servi ca sisteme de suport pentru decizii care ar putea reduce complicațiile legate de erorile medicale.

În acest context, această teză de doctorat s-a aliniat viziunii europene și prezintă rezultatele empirice ale cercetării metodelor și tehnicilor de text mining aplicate pentru dezvoltarea de sistem de suport pentru deciziile de diagnosticare a tulburării de spectru autist (TSA). Prototipul software are capacitatea de a extrage simptome specifice autismului din date text nestructurate care descriu comportamentul copiilor suspectați de TSA. În locul unei metodologii standard, orientată pe tehnologie, folosită în mod obișnuit în dezvoltarea de software, a fost urmată proiectarea centrată pe om, în care nevoile copiilor cu dizabilități, părinților și specialiștilor medicali au avut prioritate față de tehnologie. Obiectivul de a asigura o experiență plăcută și eficientă tuturor participanților la actul medical a catalizat investigarea empirică a percepției utilizatorilor cu privire la diagnosticarea TSA mediată de tehnologie, care până acum a fost puțin cercetată în literatura științifică.

1.1 STRUCTURA TEZEI DE DOCTORAT

Teza de doctorat este organizată sistematic în șapte capitole, completate de referințele bibliografice și anexe.

Capitolul 1 prezintă contextul de cercetare, contribuțiile teoretice și practice a tezei de doctorat și modalitatea în care au fost diseminate rezultatele empirice.

Capitolul 2 prezintă o analiză a stadiului actual al cunoașterii în domeniul text mining. Sunt trecute în revistă noțiuni, metode și tehnici de text mining aplicate în procesul de descoperire a cunoștințelor din date text. Accentul principal al comunității științifice a fost coeziunea text mining cu învățare automată care a determinat progresul domeniului de text mining. În acest sens, sunt prezentate principii și tehnici de învățare automată precum: clasificare, clustering, extragerea informațiilor, web mining și procesarea limbajului natural.

În Capitolul 3, accentul se îndreaptă spre aplicabilitatea text mining în domeniul medical. Este explorat potențialul text mining în facilitarea procesului de diagnosticare TSA. O revizuire a literaturii științifice din bazele de date PubMed și ACM Digital Library, a scos la iveală limitări importante a abordărilor contemporane de diagnosticare a autismului, precum erorile de diagnosticare și accesibilitatea redusă. Aplicațiile de screening care utilizează text mining ar putea servi ca instrumente valoroase de sprijin a deciziilor de diagnosticare a TSA, abordând eficient limitările actuale.

Capitolul 4 prezintă rezultatele studiului exploratoriu efectuat cu obiectivul de a examina percepția părinților și specialiștilor medicali față de diagnosticarea TSA asistată de tehnologie. Participanții la studiu au fost recrutați folosind o strategie de eșantionare neprobabilistică, conform încadrării descrise de P. Teampau [15], și au avut sarcina de a răspunde la un chestionar online. Chestionarul a evaluat variabile socio-demografice, variabile care descriu comportamentul de utilizare a tehnologiilor moderne în context TSA și variabile ce examinează percepția generală despre diagnosticarea autismului asistată de tehnologie. Analiza rezultatelor a arătat o atitudine pozitivă față de diagnosticarea TSA mediată de tehnologie și a subliniat importanța ca serviciile de sănătate să răspundă preferințelor participanților la actul medical.

Capitolul 5 prezintă metodologia de proiectare și implementare a unui prototip software capabil să identifice simptomele TSA în mod autonom, folosind text mining. Prototipul Alfa a fost dezvoltat la nivelul de maturitate tehnologică TRL 4, conform clasificării furnizate de Comisia Europeană [16]. Este explorat potențialul acestei tehnologii ca sistem de suport pentru decizii de diagnosticare a TSA, fiabil să proceseze narațiunile părinților ce descriu comportamentul copilului suspectat de autism. În plus, este analizată eficiența modelelor AI, implementate în RapidMiner, prin evaluarea comparativă a metricilor de performanță produse de algoritmii Naïve Bayes, K-Nearest Neighbors, Deep Learning și Radom Forest. Rezultatele obținute au demonstrat fezabilitatea utilizării text mining ca metodă autonomă de identificare a simptomelor TSA.

Capitolul 6 prezintă conceptualizarea testului de screening DISCOVER și metodologia de proiectare și implementare a prototipului de tehnologie Autism AI Advisor for Diagnosis (AID), dezvoltat la nivelul de maturitate tehnologică TRL 6. Prototipul Beta introduce o abordare inovatoare pentru identificarea simptomelor specifice TSA, prin folosirea text mining pentru a analiza observațiile părinților despre dezvoltarea și comportamentul copilului lor. Aplicația mobilă AID a fost implemnatată în Unity, având datele stocate în cloud, într-o bază de date NoSQL implementată folosind Firebase Realtime Database, și integrează tehnologii inteligente, precum Microsoft Azure AI Service pentru capacitățile avansate de procesare a limbajului natural. Capitolul prezintă de asemenea, un studiu empiric care a evaluat utilitatea și dezirabilitatea aplicației AID ca și sistem de suport pentru decizii de diagnosticare a TSA. Evaluarea a fost realizată printr-un experiment controlat la care au participat un număr egal de părinți cu copil cu diagnostic de autism și părinți cu copil fără diagnostic de autism. Rezultatele

studiului sugerează că text mining și tehnologiile digitale de sănătate sunt promițătoare pentru diagnosticarea timpurie a TSA.

Capitolul 7 prezintă ideile principale care se desprind din cercetările teoretice și practice efectuate, sintetizează concluziile tezei de doctorat și direcții viitoare de cercetare în text mining.

1.2 CONTRIBUȚIILE TEZEI DE DOCTORAT

Această teză de doctorat aduce mai multe contribuții teoretice și practice, după cum urmează:

1. O analiză a stadiului actual al cercetărilor în domeniu text mining, care a sintetizat în mod structurat metodele și tehnicile de preprocesare a textului, de extragere a caracteristicilor, și de învățare automată explorate în studii contemporane (Capitolul 2).

2. O examinare critică a tehnologiei în segmentul industriei medicale dedicat diagnosticării TSA, realizată printr-o revizuire sistematică a literaturii din bazele de date PubMed și ACM Digital Library folosind framework-ul PICO (Capitolul 3). Rezultatele au indicat o diversitate de abordări tehnologice pentru care s-a propus următoarea clasificare: tehnologiile medicale, biometrice, robotice, digitale și inteligente. Cercetarea a scos în evidență limitări importante și oportunitatea text mining pentru screening-ul timpuriu al autismului.

3. O analiză de piață a aplicațiilor dezvoltate pentru utilizarea în contextul diagnosticării TSA. Au fost investigate aplicațiile mobile disponibile pe Google Play (Capitolul 3). Google Play a fost ales pentru că este un magazin virtual care oferă aplicații pentru sistemul de operare Android, care este cel mai popular în rândul utilizatorilor de smartphone-uri [17].

4. O explorare a percepției și atitudinii părinților copiilor cu TSA și a specialiștilor medicali în ceea ce privește diagnosticarea tulburării de spectru autist asistată de tehnologie, realizată pe un set de date de 5712 de observații colectate de la N=56 de participanți din România (Capitolul 4). Contribuția teoretică a modelat dezvoltarea practică a tehnologiilor de diagnosticare a TSA în această teză, și a abordat o lacună notabilă în literatura științifică existentă pe acest subiect.

5. Un concept tehnologic, inovator pentru sisteme de suport pentru decizii de diagnosticare a autismului, care să identifice în mod autonom simptomele TSA în date text nestructurate ce descriu comportamentul unui copil. Prin folosirea text mining, sistemul propus îmbunătățește procesul tradițional de diagnosticare a TSA, eliminând necesitatea procesării manuale a datelor de screening. Conceptul este la nivelul de maturitate tehnologică TRL 2 (Capitolul 5).

6. Un corpus de date nou creat printr-un experiment controlat, care a implicat N=44 participanți din România. Corpusul de date conține 473 de înregistrări care surprind informații text relevante pentru evaluarea percepțiilor părinților față de comportamentul copiilor cu TSA (Capitolul 5).

7. Un prototip software proiectat și implementat la nivelul de maturitate tehnologică TRL 4. Prototipul Alfa a fost dezvoltat folosind text mining pentru crearea unui proces

autonom de clasificare cu o singură etichetă, care identifică simptomele tulburării de spectru autist în date text nestructurate. Implementarea practică a soluției tehnice propuse a fost realizată în platforma de știință a datelor RapidMiner (Capitolul 5).

8. Patru modele de inteligență artificială proiectate și implementate în RapidMiner folosind algoritmii Naïve Bayes, K-Nearest Neighbors, Deep Learning și Radom Forest pentru identificarea simptomelor TSA în date text nestructurate (Capitolul 5).

9. O analiză, bazată pe date empirice, a performanței modelelor AI antrenate cu algoritmii Naïve Bayes, K-Nearest Neighbors, Deep Learning, și Radom Forest, ce a implicat calculul metricilor: acuratețea, eroarea de clasificare, coeficientul kappa a lui Cohen și recall. Rezultatele au demonstrat că algoritmul K-Nearest Neighbors a fost cel mai performant cu o acuratețe de 78.69% (Capitolul 5).

10. Un nou test de screening TSA- DISCOVER, a fost conceptualizat și dezvoltat pentru a permite răspunsuri în format liber la întrebări. Spre deosebire de majoritatea instrumentelor de screening găsite în literatura științifică, care prezintă de obicei întrebări cu variante de răspuns, care limitează capacitatea părinților de a evalua în mod cuprinzător comportamentul copilului lor, DISCOVER depășește această limitare. Acesta permite părinților să descrie comportamentul copilului lor în limbaj natural, răspunzând la cinci întrebări redactate pe baza recomandărilor medicale de diagnosticare a TSA descrise în Manualul de Diagnostic și Statistică al Tulburărilor Mintale (Capitolul 6).

11. Un prototip software proiectat cu o abordare centrată pe om și implementat la nivelul de maturitate tehnologică TRL 6. Prototipul Beta a fost dezvoltat folosind text mining pentru crearea unui model autonom de clasificare cu etichete multiple, care identifică simptomele tulburării de spectru autist în date text nestructurate ce descriu comportamentul copiilor suspectați de autism. Implementarea practică a prototipului a fost realizată în Unity și a rezultat în aplicația Autism AI Advisor For Diagnosis, care integrează tehnologii inteligente, precum Microsoft Azure AI Service, cu capacități avansate de procesare a limbajului natural (Capitolul 6).

12. O analiză, bazată pe date empirice, a performanței modelului AI implementat în aplicația Autism AI Advisor For Diagnosis. Investigarea a implicat calculul metricilor: precizie, recall și scorul F1. Rezultatele au demonstrat o performanță bună a modelului AI, și au indicat o precizie de 91.57% (Capitolul 6).

13. Un studiu al eficienței aplicației Autism AI Advisor For Diagnosis, în condiții din lumea reală. Această cercetare a analizat comparativ rezultatele testului DISCOVER cu cele ale testului Q-CHAT (Capitolul 6).

14. O examinare a utilizabilității diagnosticării TSA asistată de aplicația Autism AI Advisor For Diagnosis, realizată printr-un experiment controlat la care au participat N=22 de părinți din România și Republica Moldova. Metricile folosite pentru evaluare au fost scalele Computer Self-Efficacy Scale și System Usability Scale, instrumentul Autism Parenting Stress Index și chestionarul de percepție a fezabilității aplicației. Cele 2156 de observații analizate din setul de date au dezvăluit o percepție pozitivă, participanții raportând un nivel înalt de mulțumire și intenția de a recomanda instrumentul de screening și altor persoane (Capitolul 6).

1.3 ARTICOLE ȘTIINȚIFICE PUBLICATE

Contribuțiile originale au fost publicate la nivel național și internațional, într-un număr total de șase lucrări științifice, dintre care cinci au fost indexate de Web of Science.

1. Mihaela Chistol, Maria-Doina Schipor, Cristina Elena Turcu. 2024. Psychological variables related to technology-mediated intervention design in autism spectrum disorder. În *Research in Developmental Disabilities*, Volum 153, 104826.
   DOI: 10.1016/J.RIDD.2024.104826
   WOSUID: WOS:001300733200001
   IF (2023): 2.9 (Q1)

2. Mihaela Chistol, Mirela Danubianu. 2024. Automated Detection of Autism Spectrum Disorder Symptoms using Text Mining and Machine Learning for Early Diagnosis. În *International Journal of Advanced Computer Science and Applications*, Volum 15, 2.
   DOI: 10.14569/IJACSA.2024.0150264
   WOSUID: INSPEC:25236964
   IF (2023): 0.7 (Q3)

3. Mihaela Chistol, Mirela Danubianu, Adina-Luminița Bărîlă. 2023. Technology-Mediated Interventions for Autism Spectrum Disorder. În *International Journal of Advanced Computer Science and Applications*, Volum 14, 12.
   DOI: 10.14569/IJACSA.2023.0141205
   WOSUID: WOS:001245055300005
   IF (2023): 0.7 (Q3)

4. Mihaela Chistol, Cristina Turcu, Mirela Danubianu. 2023. Autism Assistant: A Platform for Autism Home-Based Therapeutic Intervention. În *IEEE Access*, Volum 11, 94188-94204.
   DOI: 10.1109/ACCESS.2023.3310397
   WOSUID: WOS:001064486600001
   IF (2023): 3.4 (Q2)

5. Mihaela Chistol, Mirela Danubianu. 2021. Survey of Text Mining Research Methods and Their Innovative Applicability. În *Proceedings of the 11th International Conference The Danube - Axis of European Identity*, Galați, Romania, Volum 11, 1.

6. Mihaela Chistol. 2020. A Comparative Study of Parametric Versus Non-Parametric Text Classification Algorithms. În *Proceedings of the 2020 International Conference on Development and Application Systems (DAS)*, 208-213.
   DOI: 10.1109/DAS49615.2020.9108968
   WOSUID: WOS:000589776100039

2 STADIUL ACTUAL AL CERCETĂRILOR ÎN DOMENIUL TEXT MINING

Acest capitol prezintă o analiză a stadiului actual al dezvoltării în domeniul text mining. Acesta integrează rezultatele cercetărilor contemporane, oferind o imagine de ansamblu echilibrată atât a principiilor de bază, cât și a tehnicilor de ultimă oră care influențează traiectoria viitoare a domeniului. Primele secțiuni prezintă conceptul text mining și procesul de explorare a cunoștințelor din date text, se aprofundează metodele și tehnicile de preprocesare a textului precum: normalizare, tokenization, stemming, lematizare și se discută tehnicile de extragere a caracteristicilor. Secțiunile ulterioare prezintă principii și tehnici de învățare automată care au determinat propulsarea text mining către noi frontiere. Se discută abordările fundamentale de învățare automată care au permis calculatoarelor să imită capacitatea umană de a înțelege limbajul. În plus, este ilustrat un arbore de decizie creat de G. Miner [18] care contribuie la identificarea celor șapte tehnici de învățare automată folosite în text mining: clasificare, clustering, recuperarea informațiilor, extragerea informațiilor, extragerea conceptelor, web mining și procesarea limbajului natural.

2.1 CONCEPTUL TEXT MINING

Etimologia conceptului „text mining” își are originea în domeniul data mining și este o combinare a două concepte cheie „text” și „mining”. Text se referă la orice formă de date textuale. În general, textul este un șir de caractere scris în limbaj natural, format din cuvinte care sunt îmbinate aplicând reguli de sintaxă lingvistică. Mining este o analogie cu domeniul mineritului în care din pământ sunt extrase resurse valoroase și se referă la extragerea de informații importante din volume mari de date text. Într-un mod analog cu data mining, text mining caută să extragă informații relevante din surse de date prin identificarea și explorarea tiparelor [19].

M. Hearst [20] definește text mining ca „*descoperirea de către calculator a unor informații noi, necunoscute anterior, prin extragerea automată a informațiilor din diferite resurse scrise*”. Urmând această definiție, text mining este conceptul cercetării informatice, care transformă datele text din resursele scrise în cunoștințe. În text mining resursele scrise în principal reprezintă colecții de documente textuale semi-structurate sau nestructurate [21].

2.2 METODE ȘI TEHNICI DE TEXT MINING

Descoperirea cunoștințelor din date text este o paradigmă contemporană ce se concentrează pe explorarea computerizată a cantități mari de date text. Conceptul Knowledge Discovery in Text (KDT) a fost introdus de către R. Feldman și I. Dagan ca și framework care combină Knowledge Discovery in Databases (KDD) cu metodele de prelucrare a textului [22]. În general procesele KDD și KDT sunt considerate similare [23]. Cu toate acestea, distincția este determinată de tipul de date procesat, data mining implică date structurate, în timp ce text mining operaționalizează preponderent date nestructurate. La nivel funcțional procesul de descoperire a cunoștințelor din date text constă în parcurgerea a cinci etape ilustrate în Figura 2.1.

Figura 2.1 Procesul de descoperire a cunoștințelor din date text.

- **E_1 Colectarea datelor text**: Colectarea datelor este etapa în care se creează setul de date și care necesită un efort însemnat pentru a asigura integritatea informațiilor și a facilita accesul la acestea.
- **E_2 Preprocesarea datelor text**: Etapa de preprocesare a textului constă în aplicarea asupra eșantionului de date text a metodelor de curățare de zgomot și de date care nu aduc aport informațional. Tehnicile de preprocesare aplicate în această etapă sunt: analiza lexicală, filtrarea cuvintelor stop, normalizarea, capitalizarea, lematizarea și corectarea ortografiei [24].
- **E_3 Extragerea caracteristicilor**: Caracteristicile sunt determinate pe baza procesării conținutului documentelor. Pentru extragerea acestora textul nestructurat trebuie procesat și transformat într-un spațiu de caracteristici structurat pentru a putea aplica tehnici de modelare matematică. Metode recunoscute pentru eficiența în extragerea caracteristicilor sunt Term Frequency Inverse Document Frequency (TF-IDF) [25], Word2Vec [26] și Global Vectors for Word Representation (GloVe) [27].
- **E_4 Explorarea și descoperirea tiparelor**: Etapa de explorare constă în aplicarea algoritmilor de învățare automată în vederea descoperirii cunoștințelor din datele text. Alegerea algoritmilor aplicați este puternic influențată de scopul procesului de mining: extragerea informațiilor, clasificare, clustering sau procesarea limbajului natural. Algoritmii frecvent aplicați în această etapă sunt Naïve Bayes (NB) [28], Support Vector Machines (SVM) [29], K-Nearest Neighbors (k-NN) [30] și Deep learning [31].
- **E_5. Analiza rezultatelor**: Interpretarea rezultatelor este pasul decisiv pentru verificarea eficienței modelului generat de procesul de text mining. Practic, se evaluează abilitatea modelului de a procesa un nou set de date după ce a fost antrenat în prealabil. Evaluarea se bazează mult pe metode statistice, iar cele mai populare metrici sunt acuratețea și precizia.

2.2.1 METODE ȘI TEHNICI DE PREPROCESARE DATE TEXT

Pentru a putea fi înțelese de calculator textele trebuie simplificate. Prin urmare, etapa de preprocesare a textului devine una din cele mai importante pentru mineritul textului.

NORMALIZARE

Normalizarea este o metodă de preprocesarea a textului prin care se urmărește aducerea resursei text la un format standardizat [32]. Acest lucru îmbunătățește eficiența, deoarece contribuie la reducerea cantității de informații distincte pe care trebuie să le proceseze calculatorul. Cele mai importante tehnici de normalizare sunt normalizarea prin capitalizare (*en.*: Case Normalization) și normalizarea morfologică (*en.*: Morphological Normalization).

TOKENIZATION

V. Mohan a definit [33] tokenization ca fiind „procesul de împărțire a unui flux de conținut textual în cuvinte, termeni, simboluri sau alte elemente semnificative numite jetoane”. Urmând această definiție, tokenization este practic analiza lexicală a eșantionului de text. Conform M. Lamba și colab. [34] există două tipuri de tokenization:

1. **Word Tokenization**: este procesul de împărțire a eșantionului de text în cuvinte sau expresii de mai multe cuvinte [34].

2. **Sentence Tokenization**: este procesul de împărțire a eșantionului de text în propoziții [34].

Indiferent de tipul de tokenization, acesta este influențat de limba în care este redactat corpusul.

FILTRAREA CUVINTELOR STOP

Datele de tip text sunt formate dintr-o multitudine de cuvinte, o parte dintre aceste cuvinte nu contribuie cu informații semnificative la definirea mesajului. Aceste cuvinte sunt numite cuvinte „stop” și în această etapă de preprocesare sunt eliminate din corpusul de date text. În 1996 R. B. Myerson a propus principii de identificare a cuvintelor stop conform cărora un cuvânt este considerat stop dacă: (1) are o frecvență ridicată de apariții în text; și (2) are corelații minime cu categoriile de clasificare [35]. Filtrarea cuvintelor stop, a cuvintelor rare sau comune aduce beneficii semnificative pentru majoritatea proceselor de text mining. J. Savoy și colab. [36] au demonstrat că prin eliminarea cuvintelor stop se reduce dimensiunea de stocare a colecției de date cu 30% până la 50%.

STEMMING

J. B. Lovins [37] definește stemming ca fiind „procedura computațională care reduce toate cuvintele cu aceiași rădăcină la o formă comună”. Conform M. Lamba și colab. [34] reducerea cuvintelor flexate la forma comună este realizată prin eliminarea afixelor, care alipite la rădăcină prin prefixare și sufixare modifică sensul acestuia. Stemming este dependent de limba în care este redactat corpusul de date text, iar pentru că diferite limbi au diferite reguli lingvistice, au fost dezvoltate mai multe abordări tehnice pentru implementarea algoritmilor de stemming. Cei mai cunoscuți algoritmi de stemming sunt: algoritmul lui Lovins [37], algoritmul lui Porter [38], algoritmul lui Lancaster [39] și Snowball considerat o versiune îmbunătățită a algoritmului lui Porter [40].

LEMATIZARE

M. Lamba și colab. [34] definesc lematizarea (*en.*: lemmatization) ca procesul prin care un cuvânt este înlocuit cu forma sa din dicționar cunoscută sub numele de „lemă”. Pentru a putea realiza o astfel de conversie, un lematizator necesită un vocabular complet și metode de analiză morfologică. Un algoritm banal de lematizare este simpla căutare a cuvântului în dicționar.

2.2.2 METODE ȘI TEHNICI DE EXTRAGERE CARACTERISTICI

În procesul de explorare a textului, sursa de date poate fi reprezentată de o unitate mică precum cuvântul sau de documente voluminoase precum cărțile, paginile web sau bazele de date. Indiferent de dimensiunea sursei de date, aceasta este analizată în vederea identificării caracteristicilor reprezentative. În text mining cele mai frecvent utilizate reprezentări ale caracteristicilor sunt:

- **R_1 Reprezentare la nivel de caracter**. Reprezentarea la nivel de caractere include litere, numerali și caractere speciale. Un exemplu este Bag of Characters (BoCh) [41] o tehnică numită sac de caractere care are o aplicabilitate limitată, dat fiind faptul că nu ține cont de poziția caracterelor în eșantionul de date text.
- **R_2 Reprezentare la nivel de cuvânt**. Reprezentarea la nivel de cuvânt include caracteristici pentru cuvintele care apar în documentul text. Un exemplu este Bag of Words (BoW) [42] o tehnică numită sac de cuvinte și care folosește drept criteriu

frecvența cuvintelor în eșantionul text. Acest model ignoră complet relația semantică dintre cuvinte.

- **R_3 Reprezentare la nivel de termeni**. Reprezentarea la nivel de termeni cuprinde expresii sau cuvinte singulare care sunt extrase din eșantionul text. Metodologiile de extragere a termenilor folosesc diverse tehnici pentru extragerea și filtrarea termenilor candidați, fapt care ajută la reducerea dimensiunii și crearea unui document mai bogat din punct de vedere semantic. Cel mai frecvent utilizat fiind TF-IDF [43] care înseamnă frecvența termenului-frecvența inversă în documente. TF-IDF este o măsură statistică destinată cuantizării importanței unui cuvânt într-un document.
- **R_4 Reprezentare la nivel de concepte**. Reprezentarea caracteristicilor la nivel de concept include cuvinte, expresii cu mai multe cuvinte sau propoziții întregi care pot să nu apară în eșantionul de date text și care sunt generate manual de către un expert sau extrase prin aplicarea unor algoritmi complecși. Un exemplu este Bag of Concepts (BoC) [44] o tehnică numită sac de concepte care ține cont de aspectele de limbaj precum sinonimia și polisemia cuvintelor.

2.3 PRINCIPII ȘI TEHNICI DE ÎNVĂȚARE AUTOMATĂ ÎN TEXT MINING

J. A. Nichols [45] definește învățarea automată ca fiind „un termen general care se referă la o gamă largă de algoritmi care efectuează predicții inteligente pe baza unui set de date”. Astfel, învățarea automată este un instrument care extrage cunoștințe prin aplicarea algoritmilor asupra setului de date.

2.3.1 ABORDĂRI FUNDAMENTALE DE ÎNVĂȚARE AUTOMATĂ

În învățarea automată există trei abordări fundamentale: (1) învățare supervizată total; (2) învățare fără supervizare; și (3) învățare supervizată parțial.

1. **Învățare supervizată total**. Învățarea supervizată total este denumită și „învățare cu profesor” [46], deoarece algoritmul este antrenat pe un set de date care a fost etichetat de către un expert. Această paradigmă urmărește stabilirea unei legături între datele de intrare, numite și date de antrenare, și datele de ieșire [47]. Legăturile identificate facilitează interpretarea unor noi intrări. Astfel, această abordare este frecvent utilizată în sarcinile de clasificare.
2. **Învățare fără supervizare**. Învățarea fără supervizare nu beneficiază de cunoștințele expertului și din acest motiv este cunoscută și sub numele de „învățare fără profesor” [46]. Algoritmul trebuie să învețe independent folosind drept sursă de informare setul de date de intrare neetichetat. Această abordare este utilizată pentru sarcini de clustering și modelare de subiecte [48].
3. **Învățare supervizată parțial**. După cum sugerează numele, învățarea supervizată parțial este o abordare a învățării automate, care combină metode din învățarea supervizată total și învățarea fără supervizare [46]. Algoritmul este învățat folosind un set mic de date etichetate și un set mare de date neetichetate [34]. În această perspectivă, învățarea parțial supervizată aduce avantajul reducerii timpului de etichetare și favorizează crearea de seturi de antrenare artificiale, care sunt deosebit de utile când se urmărește creșterea volumului de date.

2.3.2 TEHNICI DE ÎNVĂȚARE AUTOMATĂ ÎN TEXT MINING

În procesul KDT etapa de explorare și descoperire a tiparelor, constă în aplicarea tehnicilor de învățare automată în vederea descoperirii cunoștințelor din datele text.

CLASIFICARE

M. Keikha și colab. [49] au definit clasificarea ca fiind „sarcina de a atribui una sau mai multe clase unui pasaj”. Urmând această definiție, clasificarea în text mining este procesul prin care unui eșantion de text îi este asociată o categorie predefinită. Procesul de învățare în clasificarea textului este unul supervizat total, deoarece este ghidat prin construirea setului de antrenare.

CLUSTERING

M. Allahyari și colab. [48] au definit clustering ca fiind „segmentarea unei colecții de documente în partiții, în care documentele din același grup (cluster) sunt mai asemănătoare cu fiecare altele decât cele din alte clustere”. Conform cadrului definit, clustering este o sarcină de NLP care grupează documentele în funcție de similitudine [34]. Termenul „cluster” denotă categorii omogene care pot fi distinse clar între ele. Identificarea clusterelor este realizată prin procesul de învățare fără supervizare, iar algoritmul nu beneficiază de date etichetate în prealabil.

EXTRAGEREA INFORMAȚIILOR

Extragerea informațiilor este procesul automat de obținere a informațiilor din date nestructurate sau semi-structurate [48], adesea provenite din surse electronice. Procesul de extragere a informațiilor vizează procesarea fiecărui document text pentru a găsi entități, relații, fapte și evenimente semnificative pentru scopul stabilit. În extragerea informațiilor nu se încearcă înțelegerea contextului documentelor, ci se definesc apriori tipurile de informații semantice care trebuie extrase din document.

WEB MINING

Web mining este tehnica dezvoltată pentru procesarea și descoperirea informațiilor utile în documentele și serviciile web [50]. Documentele web adesea sunt constituite din text organizat într-un format structurat ce conține cuvinte și termeni specifici unui limbaj de programare, hyperlink-uri și text informativ. Scopul procesului de web mining este criteriul principal după care se pot distinge trei categorii de metode de web mining: explorarea conținutului web, explorarea structurii web și explorarea utilizării web-ului [50].

PROCESAREA LIMBAJULUI NATURAL

Procesarea limbajului natural este tehnica care permite computerelor să reproducă capacitatea umană de a înțelege o limbă. Ideea ca computerele să poată înțelege limbaje și să țină conversațiile cu ființele umane, își are originea în lucrarea științifico-fantastică a lui Alan Turing publicată în 1950 [51]. Evoluția hardware a computerelor și dezvoltarea inteligenței artificiale au contribuit la implementarea în practică a ideii lui Alan Turing. În prezent, sistemele de procesare a limbajului natural sunt dezvoltate utilizând tehnici complexe, care combină învățarea automată și învățarea profundă cu metode statistice. Procesarea limbajului natural constă în parcurgerea mai multor etape secvențiale, al căror scop este să manipuleze datele text astfel încât să descopere informații de structură și conținut.

3 TEXT MINING ÎN DOMENIUL SĂNĂTĂȚII ȘI ASISTENȚEI MEDICALE

Domeniul sănătății produce cantități mari de date text în practica zilnică. Extragerea de cunoștințe din toate datele colectate, în principal nestructurate și lipsite de normalizare, este una dintre provocările majore în medicina computațională [52]. În acest sens, text mining reunește diferite tehnici pentru a obține informații valoroase din date text nestructurate, astfel încât a ajuns să fie deosebit de relevant în medicină [52].

Acest capitol, prezintă potențialul text mining în facilitarea procesului de diagnosticare a tulburării de spectru autist. Tehnicile de explorare a textului au capacitatea de a descoperi noi ipoteze și cunoștințe ascunse în date provenite din surse diferite precum interviurile, evaluările standardizate sau analizele medicale, pentru a oferi medicilor perspective importante care pot fi utilizate în luarea deciziilor informate de diagnosticare. Pentru a evalua rolul text mining în segmentul industriei medicale dedicat diagnosticării TSA, a fost efectuată o analiză a stadiului actual. În acest scop, au fost identificate lucrări relevante prin interogarea bazelor de date PubMed și ACM Digital Library, care au fost analizate cu obiectivul cercetării metodelor și tehnologiilor implicate în procesul de diagnosticare a autismului.

3.1 CONTEXT

CLINICAL TEXT MINING

Evoluția tehnologică din ultimii ani a contribuit la maturizarea algoritmilor de inteligență artificială care au facilitat dezvoltarea ariei de text mining dedicat domeniului medical numită clinical text mining. Clinical text mining se referă la metodele de procesare și extragere a cunoștințelor din textul medical [53]. Acest domeniu de cercetare combină idei și tehnici din procesarea limbajului natural, lingvistică și informatică medicală. După cum subliniază H. Dalianis [53], care a analizat aplicabilitatea text mining în domeniul clinic, text mining este folosit pentru sarcini de procesare a limbajului natural, clasificare, clustering, extragerea informațiilor, și recuperarea informațiilor.

TULBURAREA DE SPECTRU AUTIST

Tulburarea de spectru autist este o boală gravă a cărei nume își are originea în limba latină „autos”, ce înseamnă „sine”- imersiune în sine [54]. „Spectru” este un alt termen cheie care caracterizează varietatea de forme și dizabilități pe care le prezintă persoanele cu TSA [55]. Conform Manualului de Diagnostic și Clasificare Statistică al Tulburărilor Mintale (*en.:* Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)), TSA este o afecțiune neurologică și de dezvoltare complexă, caracterizată prin tulburări în interacțiunea socială și comunicare [56].

3.2 TEHNOLOGII UTILIZATE ÎN DIAGNOSTICAREA TULBURĂRII DE SPECTRU AUTIST

L. G. Wing (1981), creatoarea modelului de spectru al autismului și mama unei fete cu TSA, a spus „*Nimic nu este complet original. Toată lumea este influențată de ceea ce a fost înainte*.” [57]. Referirea la studiile anterioare este esențială pentru buna evoluție a unui domeniu de cercetare. Având ca sursă de inspirație cuvintele lui L. G. Wing, s-a efectuat o revizuire a cercetărilor publicate în segmentul industriei medicale dedicat diagnosticării TSA. S-au identificat articole relevante prin interogarea bazelor de date PubMed și ACM Digital Library. Căutarea lucrărilor științifice a fost efectuată folosind framework-ul PICO. PICO înseamnă

Pacient/Populație, Intervenție, Comparație și Rezultat și este o metodă clară de structurare a căutărilor de cercetare [58].

Datorită caracteristicilor individuale ale fiecărei baze de date, s-au adoptat diferite strategii de căutare a lucrărilor științifice. Cele 63 de lucrări științifice regăsite în urma căutării au fost evaluate folosind următoarele criterii de eligibilitate: (1) studiul abordează subiectul autismului, TSA; (2) studiul descrie metode de diagnosticare și screening; (3) cercetarea implică tehnologii moderne pentru diagnosticarea TSA; (4) studiul este disponibil complet; (5) studiul este redactat în limba engleză; (6) cercetarea este contemporană, cu restricții de perioadă cuprinsă între anii 2020 și 2024. În urma evaluării titlurilor și rezumatelor au fost eliminate un număr de 52 de articole deoarece nu îndeplineau criteriile de eligibilitate. Articolele rămase N=11 au fost citite complet, iar informațiile regăsite referitor la tehnologiile implicate în procesul de diagnosticare TSA, sunt descrise în continuare.

Tehnologii medicale: Tehnologiile medicale în diagnosticarea autismului cuprind o serie de instrumente utilizate pentru a evalua aspectele fiziologice și neurologice. Tehnici neuroimagistice precum Rezonanța Magnetică Funcțională (fMRI) [59], Tomografia cu Emisie de Pozitroni (PET) [60] și Electroencefalograma (EEG) [61] sunt utilizate pentru a studia structura și funcționarea creierului persoanelor care suferă de autism. Aceste tehnici oferă informații despre conectivitatea neuronală, modelele de activitate și anomaliile structurale asociate cu TSA.

Tehnologii biometrice: Tehnologiile biometrice presupun măsurarea și analiza caracteristicilor fiziologice ale unei persoane. În diagnosticarea TSA, tehnologiile biometrice sunt folosite pentru a evalua procesarea senzorială, recunoașterea emoțiilor și interacțiunea socială. Cele mai frecvent utilizate metode sunt sistemele de urmărire a ochilor și software-ul de recunoaștere facială.

Tehnologii robotice: Tehnologiile robotice sunt din ce în ce mai folosite în diagnosticarea autismului deoarece acestea facilitează evaluarea interacțiunii sociale și a comunicării verbale și non-verbale. Cel mai răspândit tip de roboți sunt roboții sociali cu formă humanoidă [62]. Roboții sociali sunt echipați cu senzori și capacități interactive, implică persoanele cu TSA în interacțiuni sociale și astfel oferă posibilitatea medicilor să identifice manifestările comportamentale atipice.

Tehnologii inteligente: Tehnologiile inteligente cuprind un spectru de instrumente computaționale, inclusiv AI, ML și algoritmii de DL. Aceste tehnologii revoluționează metodologiile tradiționale de diagnosticare, oferind capabilități avansate de procesare a datelor și recunoaștere a tiparelor. Prin învățarea iterativă, algoritmii AI identifică modele și corelații complexe în cadrul datelor, descoperind indicatori subtili, specifici TSA, imperceptibili pentru observatorii umani.

Tehnologii digitale: Tehnologiile digitale, cum ar fi Realitatea Virtuală (VR), Realitatea Augmentată (AR), aplicațiile mobile și dispozitivele purtabile, joacă un rol important în diagnosticarea TSA. Mediile VR și AR oferă simulări captivante care reproduc scenarii din lumea reală, facilitând observarea și evaluarea comportamentelor legate de TSA în diverse contexte. Aplicațiile mobile și dispozitivele purtabile echipate cu senzori permit monitorizarea continuă a răspunsurilor fiziologice ale indivizilor, a nivelurilor de activitate și a interacțiunilor sociale, oferind informații valoroase despre funcționarea zilnică și modelele comportamentale. Aceste tehnologii susțin de asemenea, serviciile de e-sănătate, depășind barierele geografice și crescând accesibilitatea la evaluările de diagnosticare pentru persoanele care locuiesc în zone defavorizate.

În ultimii ani, datorită introducerii tehnologiilor moderne în practica clinică pentru diagnosticarea autismului, s-au format câteva abordări promițătoare. Cu toate acestea, din cauza eterogenității TSA, prezența rezultatelor fals pozitive rămâne o provocare [63]. Literatura

științifică analizată indică câteva limitări importante a abordărilor tehnologice contemporane precum:

- Tehnologiile de screening și evaluare nu sunt întotdeauna accesibile, deoarece implică resurse financiare semnificative [64];
- Tehnologiile medicale precum fMRI, EEG, cercetarea biomarkerilor nu sunt disponibile în zonele defavorizate, implicând deplasarea pacienților pe distanțe lungi, ceea ce uneori nu este fezabil, dat fiind comorbiditățile motorii pe care le pot prezenta pacienții cu TSA;
- Tehnologiile nu produc concluzii satisfăcătoare cu privire la prezenta TSA, iar acuratețea acestora în identificarea autismului, rar are o valoare care depășește 80%;
- Antrenarea modelelor inteligente capabile să recunoască markeri specifici TSA este realizată preponderent pe eșantioane de dimensiuni mici, ceea ce limitează generalizarea rezultatelor;
- Evaluarea fezabilității tehnologiilor utilizate în diagnosticare TSA, în majoritatea cazurilor, este realizată în context demografic similar, care include factori precum cultura și locația, ceea ce poate impacta aplicabilitatea tehnologiei pe persoane cu caracteristici demografice diferite.

3.3 TEXT MINING ÎN TULBURAREA DE SPECTRU AUTIST

Determinarea precisă a trăsăturilor distinctive a persoanelor cu comportament specific TSA și diferențierea acestora de persoanelor neurotipice continuă să fie o provocare pentru specialiști în pofida progreselor înregistrate [65]. Până în prezent, nu exista o tehnologie de diagnosticare care să poată fi considerată etalon de aur, iar acest fapt crește importanța cercetării în domeniul autismului. Eforturile de cercetare trebuie îndreptate spre investigarea metodelor și tehnicilor de diagnosticare timpurie. Identificarea timpurie a TSA la copiii mici este adesea întârziată din cauza unui amalgam de factori, precum facilitățile sistemului medical, reglementări legislative, nivelul de educație și informare a membrilor familiei. Mulți părinți ai copiilor cu TSA manifestă îngrijorări cu privire la dezvoltarea copilului atunci când copilul lor are până la 12 luni [66]. Deși îngrijorările apar timpuriu, vârsta medie de diagnosticare a autismului este de 36 luni (3 ani) [66]. Statisticile arată că, în medie, 5 până la 6 luni trec de când părintele unui copil cu TSA devine îngrijorat până când se adresează unui specialist medical. Mai mult, trec peste 32 de luni până când se pune un diagnostic [67]. Astfel, se pierde o perioadă importantă de timp, în care intervenția terapeutică ar fi produs rezultate semnificative [68]. Cercetarea efectuată de O. Akinnusotu et al. semnalează că condițiile solicitante de muncă afectează starea psihologică a lucrătorilor medicali, ceea ce poate avea impact asupra diagnosticelor emise [69]. Augmentarea lucrătorilor medicali cu tehnologii moderne poate fi o soluție viabilă în combaterea burnout-ului. Cercetările contemporane indică importanța de a acorda atenție îngrijorărilor părinților și de a-i implica în procesul timpuriu de diagnosticare TSA [70]. Părinții sunt surse de încredere de informații despre dezvoltarea copiilor lor, fiind practic primii care observă manifestările comportamentale atipice. Proiectarea și implementarea tehnologiilor de screening care să încorporeze prelucrarea opiniilor părinților cu privire la dezvoltarea și comportamentul copilului ar contribui la identificarea timpurie a simptomelor autismului și ar facilita colaborarea între părinte și medic. Astfel de instrumente pot fi de asemenea, eficiente din punct de vedere al timpului de screening și din punct de vedere al costurilor de practică medicală [71], [72], [73].

Text mining oferă oportunitatea de a analiza cantități mari de date text nestructurate legate de comportamentul, dezvoltarea și interacțiunea unui copil. Algoritmii de procesare a textului pot recunoaște caracteristici specifice autismului atât în texte semi-structurate, rezultate în urma

unui screening, cât și în texte nestructurate precum narațiunile părinților. Această scalabilitate poate fi extrem de benefică medicilor, oferindu-le instrumente bazate pe tehnologia de text mining, ce pot analiza datele și semnala factori de risc și indicatori ai autismului. Prin integrarea text mining în abordările tradiționale de diagnosticare, profesioniștii din domeniul sănătății și părinții pot beneficia de următoarele avantaje:

- **Proces de screening îmbunătățit**: Permite părinților să ofere descrieri detaliate ale comportamentului copilului lor, renunțând la cadrele restrictive de răspuns comun precum sunt cele din testele convenționale;
- **Analiza automată a volume mari de date**: Încorporarea algoritmilor avansați AI, ML permite analiza automată a datelor text. Automatizarea optimizează fluxul de lucru în procesul de diagnosticare reducând semnificativ munca manuală și sporind eficiența proceselor de luare a deciziilor;
- **Accesibilitate sporită și rentabilitate:** Integrarea text mining îmbunătățește accesibilitatea, permițând capabilități de screening la distanță, accesibile din orice locație cu conexiune la internet. În plus, această abordare implică de obicei cheltuieli operaționale minime, făcând-o o soluție convenabilă din punct de vedere financiar atât pentru furnizorii de servicii medicale, cât și pentru beneficiari;
- **Diseminare facilă a rezultatelor**: Rezultatele oferite de text mining în urma procesului de screening, pot fi împărtășite rapid cu profesioniștii din domeniul medical pentru a obține o interpretare. Astfel, se promovează sinergia dintre părinte și doctor.

4 PERCEPȚIA UTILIZATORILOR FAȚĂ DE DIAGNOSTICAREA TULBURĂRII DE SPECTRU AUTIST ASISTATĂ DE TEHNOLOGIE

Comunitatea științifică a explorat noi orizonturi pentru aplicabilitatea text mining și a înțeles utilitatea acestei tehnologii în domeniul medical [74], [75], [76] în special pentru sarcini de diagnosticare [77]. Cu toate acestea, nu a acordat atenție atitudinii utilizatorilor față de procedurile de diagnosticare mediate de tehnologie. Profilul psihosocial al fiecărui utilizator joacă un rol esențial în modelarea atât a percepției asupra utilizabilității tehnologiei în practica medicală, cât și a rezultatelor obținute în urma utilizării acesteia. Din acest motiv este importantă explorarea punctului de vedere al utilizatorilor cu privire la intervențiile mediate de tehnologie. Din cunoștințele mele, acest studiu reprezintă prima explorare a percepției utilizatorilor față de diagnosticarea tulburării de spectru autist asistată de tehnologie. Astfel, scopul acestui capitol este de a raporta rezultatele obținute în urma studiului exploratoriu a opiniilor părinților și experților medicali față de diagnosticarea TSA asistată de tehnologie.

4.1 CONTEXT

Din toate afecțiunile medicale, tulburările din spectru autist prezintă unul din cele mai provocatoare domenii de aplicare ale tehnologiei în diagnostic, studiu și tratament [4]. În ciuda eforturilor considerabile în cercetările tehnologice, pentru abordarea autismului, există un gol în literatura de specialitate cu privire la atitudinea actorilor principali față de tehnologiile folosite în diagnosticarea TSA. Actorii principali sunt bineînțeles pacienții care suferă de autism, membrii familiei nucleare și specialiștii medicali. Perspectivele acestora sunt esențiale pentru a dezvolta o tehnologie de screening eficientă deoarece studiul factorilor demografici și sociali, a comportamentului de consum, a fezabilității și utilității percepută, poate furniza dovezi empirice cu privire la diverse riscuri. Înțelegerea acestor variabile poate contribui la proiectarea design-ului tehnologiilor astfel încât să fie mai accesibile, optimizate și adaptate pentru a minimiza riscurile.

4.2 EXPERIMENT

În această secțiune se explorează percepția utilizatorilor față de diagnosticarea TSA asistată de tehnologie. Prin termenul de „utilizator” mă refer la o persoană care utilizează un instrument, aplicație sau dispozitiv cu scopul de a identifica markeri asociați autismului.

Participanți: Participanții la studiu au fost recrutați folosind o strategie de eșantionare neprobabilistică, conform încadrării descrise de P. Teampau [15]. Recrutarea s-a realizat prin trimiterea de invitații prin e-mail către fundații și centre medicale din regiunea de nord-est a României, al căror domeniu de activitate este diagnosticarea și terapia copiilor cu TSA. Criteriile de includere în eșantion au fost: (1) adult cu cel puțin un copil diagnosticat cu TSA; (2) specialist medical cu experiență în diagnosticarea și îngrijirea pacienților cu TSA. Numărul de participanți care au fost incluși în cercetare a fost N=56. Participanții s-au încadrat în două grupuri reprezentative pentru actorii implicați, în situații reale, în procesul de diagnosticare TSA și anume grupul „*părinți*” (G_1) și grupul „*specialiști medicali*” (G_2). Distribuția participanților nu este egală, având un raport de 42:14, un număr de 42 (75%) dintre participanți aparținând grupului G_1 iar 14 (25%) aparținând grupului G_2.

Sarcină: Participanții la experiment au avut sarcina de a răspunde la întrebările unui chestionar folosind versiunea web a Google Forms.

Instrument: Chestionarul fost dezvoltat pentru a evalua trei domenii-cheie: variabile socio-demografice ale părintelui și specialistului medical, variabile care descriu comportamentul de utilizare a tehnologiilor moderne în context TSA și variabile ce examinează percepția generală

despre diagnosticarea autismului asistată de tehnologie. Chestionarul este organizat în trei secțiuni:

1. **Informații socio-demografice**. Participanții au furnizat informații despre mediul de reședință, categoria socio-profesională și experiența în interacțiunea cu persoanele diagnosticate cu TSA. Experiența a fost măsurată pe o scară de interval. În plus, părinții au oferit și informații despre copiii lor precum: vârsta, genul, statusul educațional, nivelul TSA, vârsta la care a fost diagnosticat și vârsta la care au apărut primele simptome ale autismului. Pentru a permite operaționalizarea variabilei „nivelul TSA", chestionarul a inclus itemii scalei M-CHAT [78].
2. **Comportamentul utilizării tehnologiilor moderne în context TSA**. Participanții au furnizat date referitoare la experiența și frecvența de utilizare a tehnologiilor moderne în context TSA.
   (a) *Frecvența utilizării tehnologiilor pe zi*, variabilă calitativă ordinală exprimată în număr de ore pe zi în care sunt utilizate tehnologii în mod regulat: „0 ore, fără utilizare", „mai puțin de 1 oră pe zi, utilizare limitată", „1-2 ore, utilizare moderată", „3-5 ore, utilizare frecventă", și „peste 5 ore, utilizare intensivă". Variabila măsoară frecvența utilizării tehnologiilor în general, indiferent de tipul acesteia.
   (b) *Frecvența utilizării tehnologiilor pe săptămână*, variabilă de raport, reprezintă numărul de zile pe săptămână în care sunt utilizate tehnologii în mod regulat: „0 zile", „1 zi", „2 zile", „3 zile", „4 zile", „5 zile", „6 zile" și „7 zile". Valoarea de zero absolut fiind dată de prima opțiune „0 zile" și semnificând lipsa de utilizare.
   (c) *Tipuri de tehnologii folosite*, variabilă calitativă nominală care reprezintă categorii de tehnologii utilizate de participant: „Aplicații pentru telefon mobil", „Aplicații pentru tabletă", „Aplicații pe calculator", „Aplicații cu realitate augmentată (AR)", „Aplicații și dispozitive pentru realitate virtuală (VR)", „Dispozitive de comunicare", „Roboți", „Tabla inteligentă", „Consolă de joc" și „Altă opțiune".
   (d) *Scopul utilizării tehnologiilor*, variabilă calitativă nominală, reprezintă contexte de utilizare a tehnologiilor, prezentată sub forma unei liste predefinite cu posibilitate de selecție multiplă. Categoriile cuprinse sunt: „Diagnostic", „Divertisment", „Educațional", „Comunicare", „Terapeutic", „Asistență", „Motricitate", „Organizatoric", „Planificare" și „Altă opțiune". Categoria „Altă opțiune" oferă posibilitatea de a specifica un scop de utilizare care nu a fost cuprins în lista de opțiuni.
   (e) *Rezultatul așteptat în urma utilizării tehnologiei,* variabilă calitativă nominală care reprezintă rezultatul așteptat de participant ca consecință a folosirii tehnologiilor: „Diagnosticarea timpurie a autismului", „Îmbunătățirea abilităților de comunicare", „Învățarea rutinelor zilnice", „Divertisment și relaxare prin joacă", „Îmbunătățirea abilităților sociale", „Gestionare eficientă a activităților prin utilizarea planificatorului", „Îmbunătățirea abilităților motrice", „Învățarea exprimării și înțelegerii sentimentelor/emoțiilor", „Îmbunătățirea abilităților de scriere și citire" și „Altă opțiune".
   (f) *Costul de achiziționare a tehnologiei,* variabilă calitativă ordinală, opțiunile de răspuns au fost organizate în categorii de prețuri ordonate: „0 lei", „Între 1 și 20 de

lei", „Între 20 și 50 de lei", „Între 100 și 200 de lei", „Între 200 și 500 de lei", „Între 500 și 1000 de lei", „Între 1000 și 2000 de lei" și „Peste 2000 de lei". Categoriile de preț au fost raportate la salariul minim net în Romania în 2023 și la costurile tehnologiilor pe piață.

3. **Percepția utilizatorilor asupra diagnosticării TSA asistată de tehnologie**. Participanții au furnizat informații referitor la percepția utilizabilității tehnologiilor pentru diagnosticarea TSA.
   (a) *Importanța diagnosticării timpurii a autismului*, variabilă calitativă textuală, permite participantului să ofere un răspuns liber și să detalieze percepția asupra diagnosticării timpurii a tulburărilor din spectru autist.
   (b) *Importanța tehnologiilor de diagnosticarea TSA*, variabilă calitativă ordinală, evaluată pe o scală Likert în 5 puncte a importanței percepute referitor la tehnologiile de diagnosticare, de la „1, foarte mică" la „5, foarte mare".
   (c) *Experiența în utilizarea instrumentelor tehnologice în vederea diagnosticării TSA*, variabilă calitativă nominală care măsoară gradul de experiență a participanților în ceea ce privește utilizarea instrumentelor tehnologice pentru diagnosticarea TSA.
   (d) *Fezabilitatea percepută*, variabilă calitativă textuală, reprezintă viabilitatea percepută de participanți referitor la tehnologiile de diagnosticare TSA.
   (e) *Experiența emoțională*, variabilă calitativă nominală, capturează stări emoționale ale participanților.

4.3 REZULTATE

Rezultatele acestui studiu au fost analizate comparativ pentru a distinge între percepțiile părinților care au copii cu diagnostic de TSA (G_1) și percepțiile specialiștilor medicali calificați în diagnosticarea TSA (G_2).

REPREZENTATIVITATEA EȘANTIONULUI DE PARTICIPANȚI

În Figura 4.1 și 4.2 sunt prezentate rezultatele analizei datelor socio-demografice care arată că participanții din mediul urban au avut o reprezentativitate mai mare atât în grupul G_1 (61.90%) cât și în G_2 (85.71%). Numărul participanților care au reședința în mediul rural este mai mic, în G_1 (38.08%) și în G_2 (14.28%), însă suficient pentru a surprinde și perspective specifice acestui habitat, care este considerat defavorizat comparativ cu mediul urban [79].

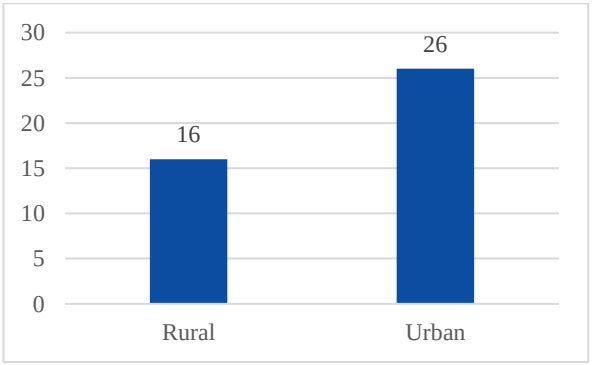


Figura 4.1 Distribuția pe mediul socio-demografic a participanților la experiment din grupul G_1.

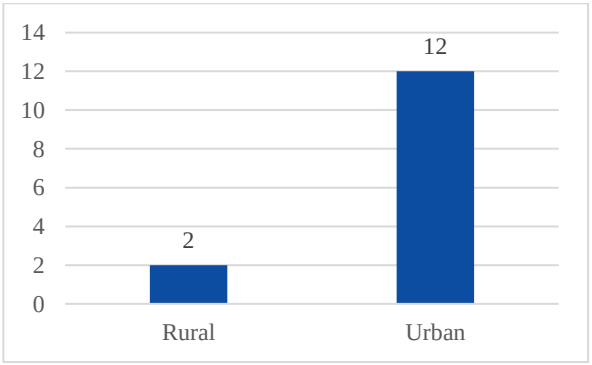


Figura 4.2 Distribuția pe mediul socio-demografic a participanților la experiment din grupul G_2.

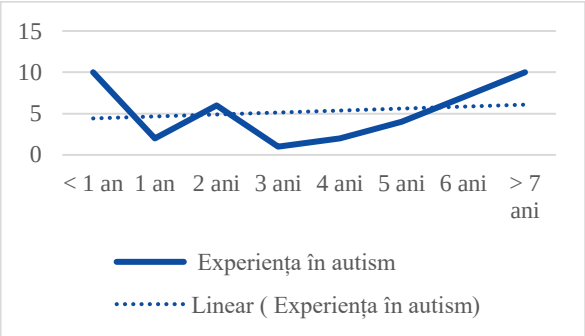


Figura 4.3 Experiența în interacțiunea cu persoanele diagnosticate cu TSA a participanților la experiment din grupul G_1.

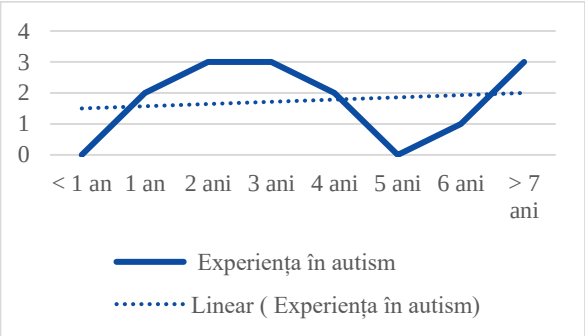


Figura 4.4 Experiența în interacțiunea cu persoanele diagnosticate cu TSA a participanților la experiment din grupul G_2.

În ceea ce privește experiența în TSA, eșantionul demonstrează o reprezentativitate mai mare pentru participanții experimentați și pentru cei cu o experiență limitată, după cum se poate observa în Figura 4.3 și Figura 4.4. Distribuția inegală a experienței este favorabilă acestei cercetări deoarece astfel au fost integrate opinii ale părinților care recent au trecut printr-un proces de diagnosticare a copilului și puncte de vedere ale experților medicali cu cunoștințe aprofundate despre autism și metodele de diagnosticare.

Participanții din G_1 au raportat și date demografice care caracterizează copiii lor ce sunt diagnosticați cu autism. Copiii aveau vârste cuprinse între 2 ani și 19 ani. Raportul pe genuri este de 33:9 cu o reprezentativitate mai mare pentru copiii de genul masculin 78.57% față de copiii cu genul feminin 21.42%. Evoluția primelor simptome ale autismului la copii și evoluția diagnosticării TSA este prezentată în Figura 4.5, axa X (orizontală) reprezintă timpul în luni, iar axa Y (verticală) numărul cumulat de copii. Graficul arată perioada de timp dintre primele simptome și momentul în care copilul a fost diagnosticat (*en.*: Time To Diagnosis (TTD)), care a fost în medie de 9.74 de luni. Majoritatea participanților din G_1 (76.19%) au considerat că un diagnostic timpuriu ar fi avut un impact pozitiv asupra dezvoltării copilului.

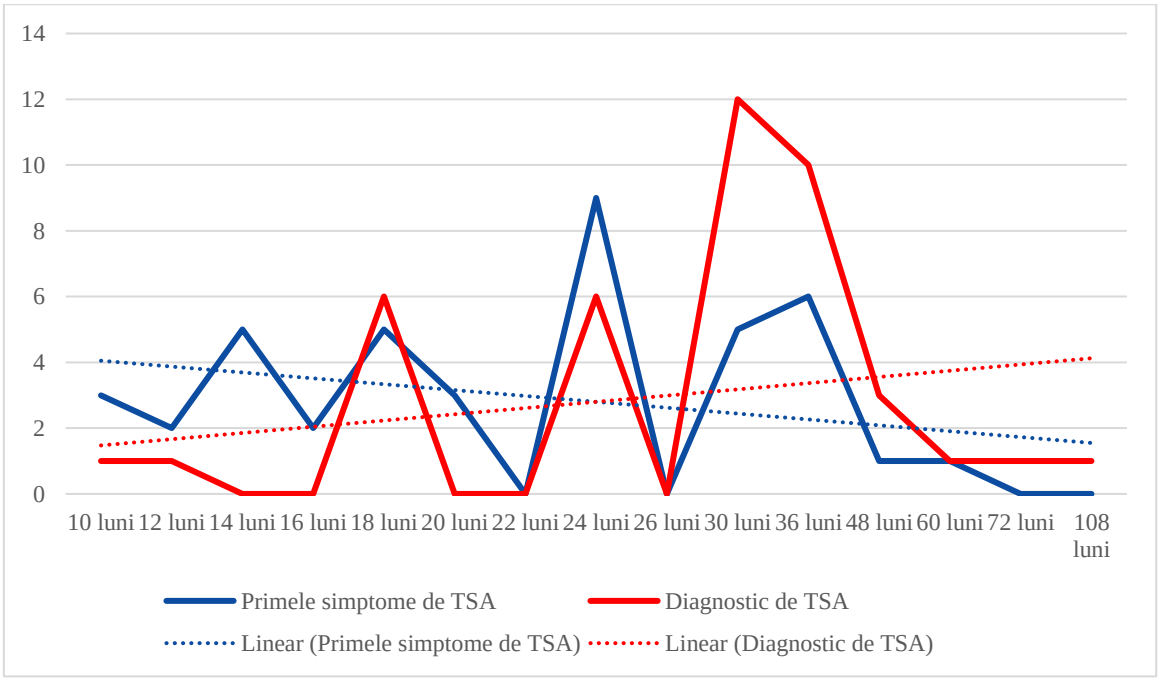


Figura 4.5 Grafic cumulativ ce prezintă evoluția primelor simptome TSA comparativ cu evoluția diagnosticelor TSA la copiii participanților din grupul G_1.

COMPORTAMENTUL DE UTILIZARE A TEHNOLOGIILOR ÎN CONTEXT TSA

Comportamentul de utilizare a tehnologiilor în context TSA a fost examinat din perspectiva mai multor variabile auto-raportate de participanți precum: timpul petrecut în utilizarea tehnologiilor, scopul pentru care sunt utilizate și efectele scontate. Analiza tendințelor de utilizare a arătat că majoritatea participanților din G_1 de 19 (45.23%) nu folosesc tehnologii, în timp ce restul au raportat „utilizare limitată" de mai puțin de 1 oră pe zi, 11 (26.19%), „utilizare moderată" 9 (21.42%), „utilizare frecventă" 1 (2.38%) și „utilizare intensivă" de peste 5 ore pe zi la 2 (4.76%). Specialiștii medicali din grupul G_2 au raportat o frecvență de utilizare diferită, nu a existat un comportament care să fie majoritar, 6 (42.85%) au spus că nu folosesc tehnologii și 6 (42.85%) au declarat „utilizare limitată", restul de 2 (14.28%) precizând „utilizare moderată".

Timpul petrecut în medie de un participant folosind tehnologii este de 1 zi pe săptămână pentru G_1, ceea ce este cu 0.22 de zile mai mult decât timpul raportat de participanții din G_2 de 0.78 zile pe săptămână. Cele mai utilizate tehnologii în ambele grupuri G_1 și G_2 au fost aplicațiile mobile (47.61% în G_1 și 28.57% în G_2), urmate de aplicațiile pentru tabletă G_1 (16.66%) și G_2 (21.42%). Analiza tehnologiilor cu realitate augmentată sau virtuală arată că părinții din G_1 nu au folosit deloc AR și VR în timp ce specialiștii medicali din G_2 au experimentat cu tehnologii AR în proporție de 7,14%. Când vine vorba de aplicații pentru calculator (9,52%), dispozitive (14.25%), console de joc (4.76%) și roboți (4.76%), participanții din G_1 au manifestat un interes sporit față de cei din G_2 (0%).

În cadrul studiului exploratoriu s-a cercetat și contextul de utilizare a tehnologiilor. Participanții au ales dintr-o listă predefinită, cu posibilitate de selecție multiplă, contextele în care au folosit tehnologii. Contextul educațional a fost raportat de majoritatea, constituind 59.52% dintre participanții din G_1 și 42.85% din G_2. Divertismentul a fost al doilea cadru preferat de ambele grupuri cu 23 (54.76%) din G_1 și 5(35.71%) din G_2 urmat de comunicare G_1 13 (30.95%) și G_2 4 (28.57%). Părinții au folosit mai frecvent tehnologiile cu scop terapeutic 13 (30.95%) față de cadrele medicale 3 (21.42%). Tendința poate fi explicată prin rolul semnificativ pe care părintele îl joacă ca principală sursă de sprijin al copilului său care suferă de autism. Contextul de motricitate asociat cu activitatea și reabilitarea fizică a fost indicat într-un procent mai mic G_1 7 (16.66%) și G_2 1 (7.14%). Scopurile de asistență și planificare au fost cel mai puțin raportate cu doar 1 (2.38%) participant din grupul G_1. Rezultatele aferente contextului de diagnosticare a autismului a pus în evidență un aspect de interes, mai exact 5 (11.90%) părinți din grupul G_1 și 4 (28.57%) cadre medicale din grupul G_2 au explorat tehnologii pentru scopul de diagnosticare TSA. Această constatare arată interesul unor participanți de a experimenta cu abordări noi, diferite de practicile tradiționale de diagnosticare a autismului și în același timp prudența celorlalți participanți față de schimbări.

Tehnologiile au potențialul de a îmbunătăți calitatea vieții [80]. Participanții la studiu au înțeles acest aspect și astfel au avut așteptări mari cu privire la efectele pozitive în urma utilizării tehnologiei: îmbunătățirea abilităților de comunicare, îmbunătățirea abilităților sociale, gestionare eficientă a activităților prin utilizarea planificatorului și diagnosticarea timpurie a autismului au fost cel mai frecvent raportate.

PERCEPȚIA UTILIZATORILOR ASUPRA DIAGNOSTICĂRII TSA ASISTATĂ DE TEHNOLOGIE

Importanța atribuită de eșantion pentru tehnologiile de diagnosticare a autismului este de o semnificație majoră. Cei mai mulți dintre specialiștii medicali 5 (35.71%) din grupul G_2 au evaluat importanța pe o scală Likert în 5 puncte la „5, foarte mare" fiind susținuți în opinie de

13 (30.95%) părinți din G_1. Participanții au considerat de asemenea, că tehnologiile de screening sunt de o importanță „4, mare" în procent de 21.42% de participanți din G_1 și 21.42% din G_2. Doar 2 (4.76%) părinți și 2 (14.28%) specialiști medicali au raportat o importanță „1, foarte mică". Importanța tehnologiilor de diagnosticare TSA este strâns corelată de importanța percepută de participanți pentru diagnosticare TSA în sine.

Fezabilitatea aplicațiilor software ca suport în diagnosticarea autismului a fost evaluată printr-o analiză calitativă a răspunsurilor textuale ale participanților la întrebarea: "*Care este opinia dumneavoastră referitor la utilizarea aplicațiilor pentru diagnosticarea timpurie a autismului?".* Această abordare a urmărit să surprindă și să interpreteze perspectivele participanților cu privire la caracterul practic al instrumentelor de diagnosticare. Dat fiind contextul profesional medical, specialiștii din G_2 au considerat viabile practicile de diagnosticare mediate de tehnologie. Răspunsul participantul P_{14} este prezentat în Figura 4.6 și exprimă opinia că tehnologiile pot spori accesibilitatea serviciilor medicale din perspectiva resurselor financiare necesare pentru diagnosticarea TSA și a timpului de screening.

> *„Pot oferi întregii familii o accesare mai ușoară și pot trece barierele ce țin de: cost, timp, amplasare geografică, servicii limitate și lipsă de informare."*

Figura 4.6 Percepția participantului P_{14} asupra fezabilității aplicațiilor software de diagnosticare TSA.

> *„Cred că este în folosul familiei ajutându-i să conștientizeze faptul că există probleme, mulți părinți trecând prin faza de negare."*

Figura 4.7 Percepția participantului P_{42} asupra fezabilității aplicațiilor software de diagnosticare TSA.

> *„Vine în ajutorul părinților care își pun multe semne de întrebare și nu au curajul de a merge la un medic specializat."*

Figura 4.8 Percepția participantului P_{48} asupra fezabilității aplicațiilor software de diagnosticare TSA.

Figura 4.7 și Figura 4.8 ilustrează perspectivele participanților P_{42} și P_{48} din grupul G_1. Părinții au considerat că aplicațiile de diagnosticare TSA pot funcționa ca sisteme de validare a simptomelor autismului înaintea unui diagnostic oficial primit de la medic. În perspectiva acestora, aplicația va aborda și valida preocupările lor cu privire la comportamentul copilului, încurajându-i să investigheze în continuare starea de sănătate a copilului.

Ultima variabilă analizată este experiența emoțională, variabilă specifică grupului G_1, al părinților, deoarece aceștia sunt supuși stresului atunci când parcurg un proces de diagnosticare a copilului suspectat de TSA. Rezultatele arată că majoritatea participanților au considerat că utilizarea aplicațiilor în detectarea simptomelor specifice autismului, poate fi mai confortabilă pentru părinți din punct de vedere emoțional.

5 TEXT MINING PENTRU DIAGNOSTICAREA TULBURĂRII DE SPECTRU AUTIST

Capitolul anterior a explorat în detaliu atitudinile utilizatorilor față de diagnosticarea TSA asistată de tehnologii moderne. Rezultatele obținute au încurajat explorarea text mining ca și sistem de suport pentru decizii de diagnosticare, capabil de a identifica tipare în date text ce pot sugera prezența simptomelor TSA la populația examinată.

În acest capitol, se prezintă metodologia de cercetare pentru proiectarea și implementarea, în RapidMiner, a unui prototip autonom de identificare a simptomelor TSA. Se explică etapele de instruire ale modelului de inteligență artificială și se detaliază metodele și tehnicile de text mining utilizate pentru a obține o eficiență superioară. În plus, se analizează performanța modelului prin evaluarea comparativă a acurateței, erorii de clasificare, recall și coeficientul kappa al lui Cohen, produse de modelul antrenat folosind algoritmii Naïve Bayes, K-Nearest Neighbors, Deep Learning și Radom Forest.

5.1 CONTEXT

Progresele în cercetare și învățare automată sugerează că sistemele de suport pentru deciziile clinice pot fi un instrument inovator și eficient pentru susținerea progreselor în domeniul sănătății ce vizează diagnosticarea și îngrijirea copiilor cu diverse neurodezvoltări, inclusiv TSA [81].

SISTEM DE SUPORT PENTRU DECIZII CLINICE

Un sistem de suport pentru deciziile clinice (*en.*: clinical decision support system (CDSS)) este mulțimea instrumentelor computerizate și necomputerizate care oferă personalului medical și pacienților informații specifice cu privire la starea de sănătate și recomandări de îngrijire [82]. Conform A. T. M. Wasylewicz și colab. CDSS pot fi împărțite în dependență de funcție în două categorii: (1) sisteme care răspund la întrebarea „*Ce este adevărat?*”; (2) sisteme care răspund la întrebarea „*Ce să fac?*”, ambele întrebări se referă la pacient [82]. În categoria (1) sunt incluse sistemele de suport pentru decizii de diagnosticare (*en.*: diagnosis decision support system (DDSS)), iar în (2) sistemele de suport pentru management.

MODEL AI DE PREDICȚIE CLINICĂ

Un model de AI este un program informatic capabil să imită raționamentul uman datorită instruirii sale folosind algoritmi de inteligență artificială [83]. Modelul este antrenat pe un set de date reprezentativ pentru sarcina propusă aplicând metode statistice care conduc la stabilirea unor corelații dintre variabilele de intrare și variabilele de ieșire. Variabilele de intrare independente sunt numite caracteristici, iar variabilele de ieșire sau dependente sunt numite etichete sau clase [34]. În învățarea supervizată un model de predicție este modelul capabil să prezică cu acuratețe etichetele pentru date noi. Similar un model de predicție clinică încearcă să prezică șansele unui pacient de a avea un anumit rezultat pe baza datelor sale clinice [84].

5.2 METODOLOGIE DE LUCRU

Diagnosticarea TSA este o sarcină dificilă deoarece etiologia și factorii care determină apariția autismului sunt necunoscuți. În plus, spectrul larg al simptomelor și lipsa unui test medical precis, precum cel de sânge, îngreunează procesul de diagnosticare. Procesul tradițional de diagnosticare prezentat în Figura 5.1 implică factorul uman, reprezentat de medic, care analizează parametrii pacientului și aplică strategii psihologice de observație pentru a iden-

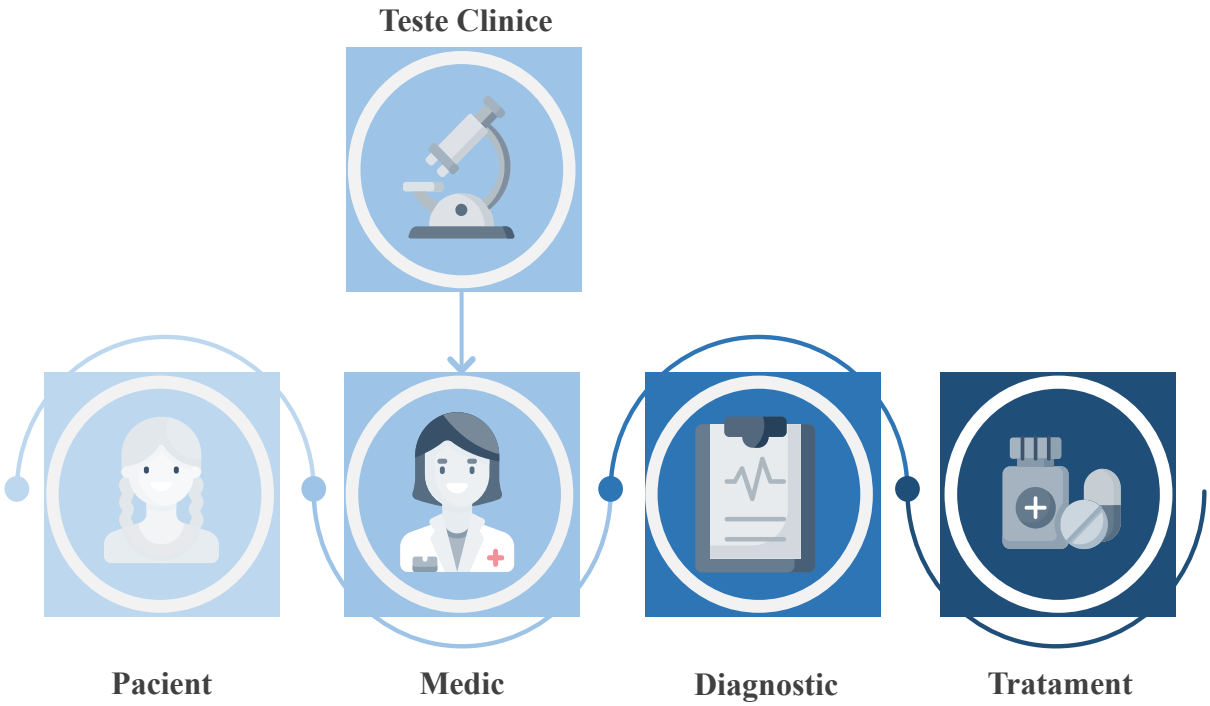


Figura 5.1 Proces tradițional de diagnosticare TSA.

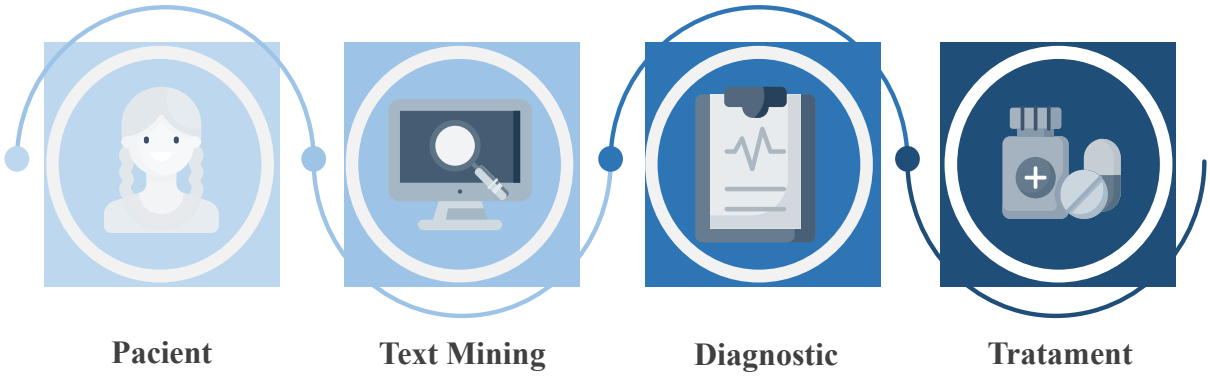


Figura 5.2 Proces autonom de diagnosticare TSA.

tifica simptome ale TSA. Cercetarea efectuată de O. Akinnusotu și colab. [69] semnalează că condițiile solicitante de muncă afectează starea psihologică a lucrătorilor medicali, ceea ce poate avea impact asupra diagnosticelor și tratamentelor emise. Augmentarea lucrătorilor medicali cu tehnologii moderne precum DDSS poate fi o soluție viabilă în combaterea burnout-ului. Figura 5.2 prezintă conceptul tehnologic propus pentru procesul autonom de diagnosticare TSA care elimină factorul uman și introduce text mining și algoritmi de învățare automată cu scopul de a descoperi simptome ale autismului în datele pacientului.

Detectarea simptomelor TSA în date text nestructurate implică o serie de pași complecși pentru a extrage informații relevante și pentru a construi un model capabil să identifice simptome asociate cu autismul. Etapele procesului sunt inspirate din metodologii consacrate de cercetare precum KDT:

- **E_1 Preprocesarea datelor**: Colectarea și preprocesarea datelor este o etapă importantă în procesul identificării simptomelor autismului și trebuie realizată în colaborare cu

cadre medicale calificate să emită concluzii asupra stării de sănătate a participanților, concluzii care să se reflecte în etichetele asociate datelor text;

- **E_2 Antrenarea modelului**. În această etapă se antrenează un model de învățare automată, folosind drept set de instruire datele etichetate la pasul anterior, și se aplică algoritmi pentru clasificarea textului;
- **E_3 Testarea modelului**. Etapa de testare a modelului constă în furnizarea unui set de date de testare către modelul antrenat și calcularea unor indicatori de performanță;
- **E_4 Analiza rezultatelor**. Analiza rezultatelor constă în evaluarea performanței modelului antrenat, folosind indicatorii calculați la pasul anterior precum acuratețea și eroarea de clasificare. În dependență de rezultatele evaluării se pot ajusta parametrii modelului și se pot explora noi algoritmi de învățare automată.

Procesul autonom de identificare a simptomelor TSA în date text nestructurate este unul empiric și trebuie rafinat iterativ, experimentând cu reprezentarea caracteristicilor, ajustarea pașilor de preprocesare și algoritmii de învățare automată aplicați.

5.3 EXPERIMENT

Diagnosticarea timpurie a autismului este esențială pentru optimizarea rezultatelor terapeutice și creșterea eficacității procesului de intervenție. Astfel, vârsta precoce a copiilor și abilitățile reduse de scriere și citire, în această etapă a vieții, au determinat implicarea adulților în experiment. Din întreg spectrul populației cercetarea a vizat un eșantion specific. Eșantionarea a fost realizată pe baza unei strategii neprobabilistice (nealeatoare), iar criteriile de includere în eșantion au fost: (1) adult cu cel puțin un copil diagnosticat cu TSA; (2) nivel C_2 pentru limba română conform Cadrul European Comun de Referință pentru Limbi (*en.*: Common European Framework of Reference (CEFR)) [85].

Participanți: Un total de N=44 de participanți din România au luat parte în mod voluntar la experiment. Pentru a surprinde o diversitate a experiențelor personale și a statutelor socioeconomice, studiul a inclus participanți din diferite organizații administrative urbane și rurale. Datele arată că procentul participanților care locuiesc în zone rurale este de 38.64% (27), procent mai mic comparativ cu procentul participanților din zonele urbane 61.36% (17). Persoanele de genul feminin și genul masculin sunt diagnosticate cu autism în mod disproporționat [86]. Motivele principale fiind manifestarea simptomelor TSA cu o intensitate mai scăzută la persoanele de genul feminin [86]. Raportul între genurile copiilor participanților la experiment care este de 35:9, 79.54% sunt băieți și 20.45% sunt fete.

Sarcină: Participanții la experiment au avut sarcina de a răspunde la întrebările unui chestionar folosind versiunea web a Google Forms.

Instrument: Chestionarul fost dezvoltat pentru a evalua trei domenii-cheie: variabile demografice părinte, variabile demografice copil diagnosticat cu TSA și variabile relevante pentru evaluarea percepțiilor părintelui față de comportamentul copilului. Dimensiunea părinte s-a concentrat pe experiența acestuia în contextul TSA. Dimensiunea copil a inclus nivelul de severitate TSA care a fost operaționalizat cu ajutorul testului M-CHAT [78]. A treia dimensiune a examinat evaluarea părintelui privind comportamentul copilului, prin întrebări cu răspuns liber care au facilitat expunerea completă a opiniilor. Răspunsurile participanților la chestionar au fost citite complet și curățate de înregistrări incomplete, apoi au fost salvate într-un fișier Excel care conține 473 de înregistrări.

CREARE SET DE DATE TEXT

Datele brute colectate de la participanți au fost analizate și etichetate împreună cu un medic

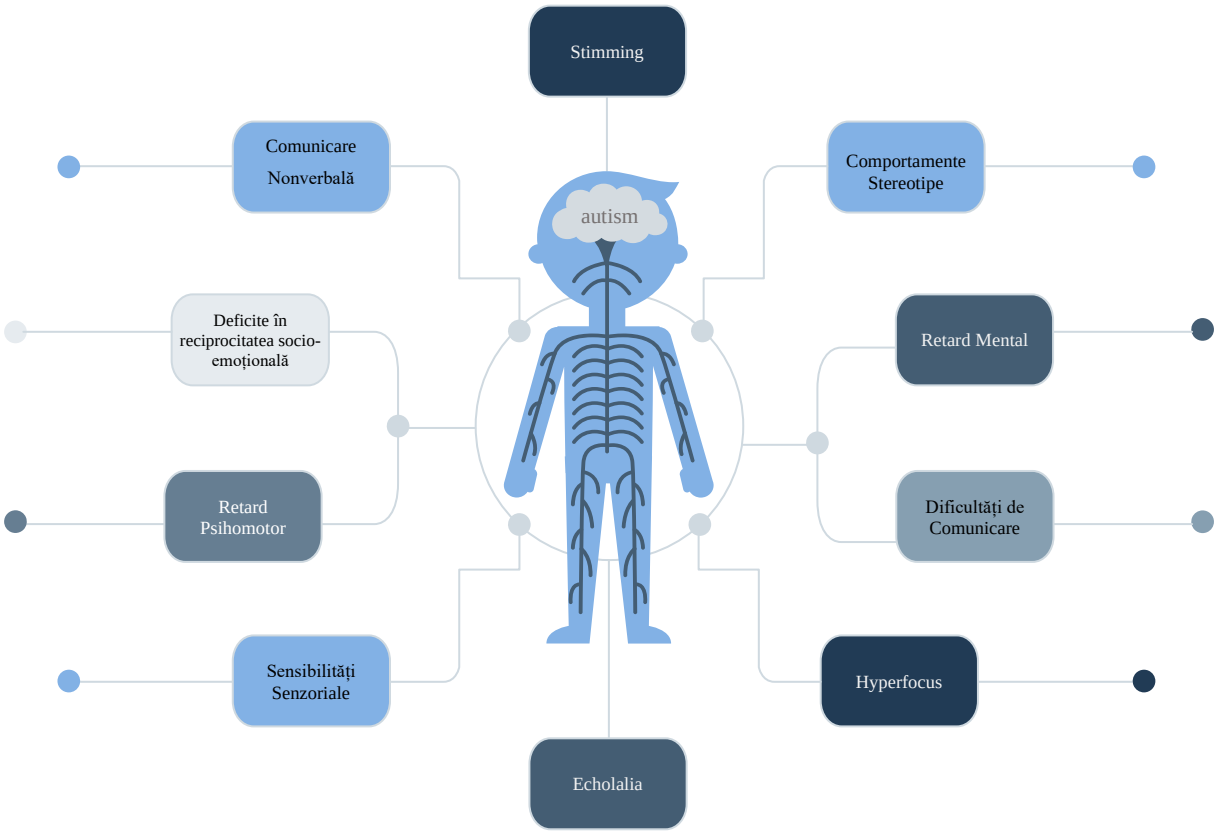


Figura 5.3 Simptome comune pe care le prezintă frecvent persoanele cu tulburare din spectrul autist.

specializat în tratarea TSA. Schema de etichetare conține 19 etichete, dintre care 18 etichete reprezintă simptome specifice autismului și o etichetă specială „Asymptomatic" care semnifică lipsa simptomelor. Etichetele au fost extrase din descrierea medicală a diagnosticului autismului redactată în DSM-5 [87]. Datele etichetate au fost traduse din limba română în limba engleză, pentru a asigura compatibilitatea cu instrumente de procesare a limbajului natural și a facilita diseminarea rezultatelor în literatura științifică.

Fiecare copil cu TSA are un model comportamental complex, însoțit de manifestări asociate cu diferite grade de severitate. În unele cazuri, copii afișează simptome ale tulburării de spectru autist chiar de la începutul dezvoltării lor, în timp ce alții pot evolua în mod tipic în primele luni sau chiar ani, iar mai târziul își pierd abilitățile lingvistice dobândite anterior și devin retrași social. Simptomele apar de obicei până la vârsta de 2 ani [88]. Figura 5.3 prezintă simptomele comune manifestate de copii cu TSA: deficite în reciprocitatea socio-emoțională, dificultăți de comunicare și comunicare nonverbală, comportamente stereotipe, diverse sensibilități senzoriale, Stimming, Echolalia, Hyperfocus, retard mental și retard psihomotor. Simptomele comune pe care le prezintă persoanele cu TSA aparțin celor cinci categorii de manifestări A, B, C, D și E descrise în ghidul de diagnosticare a autismului [81].

PREPROCESARE DATE TEXT

Prototipul autonom de identificare a simptomelor autismului a fost implementat în RapidMiner, fiind cel mai popular software pe piața de data mining [89]. S-a folosit pachetul Text Processing pentru preprocesarea setului de date, deoarece datele conțineau cuvinte care nu aduceau un aport informativ și care măreau dimensiunea vocabularului contribuind astfel la ridicarea complexității de calcul. Preprocesarea textului este o etapă crucială în text mining, care implică curățarea și transformarea datelor text brute într-un format adecvat pentru analiză. Metodele (M) de preprocesare a textului aplicate au fost:

- **M1** Capitalizarea;
- **M2** Analiza lexicală;
- **M3** Filtrarea cuvintelor stop, comune și rare;
- **M4** Stemming;
- **M5** n-Grams;
- **M6** Determinarea ponderilor.

M1 Capitalizarea este procesul prin care toate caracterele din text sunt transformate în majuscule sau minuscule. În procesul de identificare a simptomelor autismului datele text au fost transformate în minuscule, adică litere mici (Exemplu 5.1).

Exemplu 5.1

Question: *Describe your concerns about your child's behavior.*
Response: *He does not interact with the other children.*

După aplicarea capitalizării textul va fi astfel:
Question: *describe your concerns about your child's behavior.*
Response: *he does not interact with the other children.*

M2 Analiza lexicală sau Tokenization este o tehnică de preprocesare prin care textul este împărțit în elemente semnificative precum: simboluri, cuvinte sau fraze, numite jetoane [90]. Exemplu 5.2 prezintă rezultatul aplicării analizei lexicale asupra unei înregistrări din setul de date.

Exemplu 5.2

Question: *describe your concerns about your child's behavior.*
Response: *he is non-verbal.*

După aplicarea analizei lexicale jetoanele sunt următoarele:
Question: *{„describe"; „your"; „concerns"; „about"; „your"; „child"; „s"; „behavior"}*
Response: *{„he"; „is"; „non"; „verbal"}*

M3 Filtrarea cuvintelor stop, comune și rare este o metodă de preprocesare prin care din text sunt eliminate cuvinte care se încadrează în categoria cuvintelor stop, comune sau rare. În procesul de identificare a simptomelor autismului s-a utilizat operatorul „Filter Stopwords", care are încorporat lista de cuvinte stop pentru limba engleză și conține cuvinte precum „at", „etc", „if", „or" etc. În Exemplu 5.3 se poate observa rezultatul filtrării cuvintelor stop dintr-un răspuns al părintelui care descrie îngrijorările referitoare la comportamentul copilului. Din textul inițial au fost eliminate cuvinte precum „am", „the", „of", „that", „about", „he", „has", „when", „not" și „always" păstrând nealterat sensul mesajului. Filtrarea cuvintelor aduce beneficii semnificative pentru reducerea complexității de calcul pentru algoritmi de învățare automată dependenți de numărul de caracteristici.

Exemplu 5.3

Question: *{„describe"; „your"; „concerns"; „about"; „your"; „child"; „s"; „behavior"}*
Response: *{„i"; „am"; „worried"; „about"; „the"; „fact"; „that"; „he"; „has"; „various"; „fears"; „that"; „he"; „gets"; „angry"; „quickly"; „when"; „things"; „are"; „not"; „like"; „hi"; „wants"; „the"; „fact"; „that"; „he"; „is"; „not"; „always"; „careful"; „that"; „he"; „is"; „not"; „aware"; „of"; „the"; „danger"}*

După filtrarea cuvintelor stop rezultatul este următorul:
Question: *{„describe"; „concerns"; „child"; „s"; „behavior"}*
Response: *{„i"; „worried"; „fact"; „various"; „fears"; „gets"; „angry"; „quickly"; „things"; „hi"; „wants"; „fact"; „careful"; „aware"; „danger"}*

M4 Stemming. În morfologia lingvistică stemmin-ul este procesul de reducere a cuvintelor flexate, derivate la forma lor de bază, la rădăcină [91]. Rădăcina este partea a unui cuvânt care

este comună tuturor variantelor sale flexate. Procesul de stemming implică eliminarea prefixelor sau sufixelor. Pentru a realiza acest lucru s-a folosit algoritmul de stemming a lui Porter. Algoritmul de stemming a lui Porter elimină sufixele dintr-un cuvânt din limba Engleză pentru a obține rădăcina acestuia [92].

M_5 n-Grams este o metodă de preprocesare a textului folosită preponderent pentru extragerea caracteristicilor. Un n-Gram poate fi definit ca o serie de jetoane consecutive de lungime N. În procesul de identificare a simptomelor autismului s-a utilizat n-Grams pentru surprinderea relațiilor dintre cuvinte. Exemplu 5.4 prezintă aplicarea metodei pentru generarea bigrams cu $N = 2$ pe textul rezultat din etapa de stemming.

Exemplu 5.4

Question: *{„child”; „make”; „sentenc”; „word”}*
Response: *{„child”; „form”; „multi”; „word”; „sentenc”; „help”}*

După generarea n-Grams rezultatul este următorul:
Question: *{„child”; „child_make”; „make”; „make_sentenc”; „sentenc”; „sentenc_word”; „word”;}*
Response: *{„child”; „child_form”; „form”; „form_multi”; „multi”; „multi_word”; „word”; „word_sentenc”; „sentenc”; „sentenc_help”; „help”}*

M_6 Determinarea ponderilor este procesul prin care este cuantificată importanța caracteristicilor în setul de date text. Fiecărei caracteristici îi este asociată o valoare numită pondere, care semnifică cât de indispensabilă este aceasta pentru procesul de text mining. În procesul de identificare a simptomelor autismului s-a folosit TF-IDF pentru determinarea ponderilor.

APLICARE ALGORITMI DE ÎNVĂȚARE AUTOMATĂ

Procesul de învățare automată pentru identificarea simptomelor autismului din date text a implicat explorarea a patru algoritmi consacrați pentru sarcinile de clasificare în ML:

- **Naïve Bayes**. Clasificatorul Naïve Bayes este un algoritm de învățare supervizată bazat pe teorema lui Bayes. Algoritmul calculează probabilitatea fiecărei clase (etichete) și apoi alege clasa cu cea mai pare probabilitate [93].
- **K-Nearest Neighbors**. k-NN este un algoritm de învățare supervizată robust și fezabil pentru problemele de clasificare. Algoritmul k-NN funcționează prin găsirea celor mai apropiați k vecini de un anumit punct de date, pe baza unei metrici cum este distanța euclidiană. Clasa sau valoarea punctului de date este apoi determinată de media celor k vecini [94].
- **Deep Learning**. Algoritmul de Deep Learning se bazează pe o rețea neuronală artificială cu mai multe straturi, care este antrenată cu coborâre în gradient stocastic folosind propagarea înapoi. Fiecare strat constând din neuroni care folosesc diverse funcții de activare precum tanh, redresor și maxout [95].
- **Radom Forest**. Random Forest este un algoritm de învățare automată care utilizează un ansamblu de arbori de decizie pentru a face predicții. Fiecare arbore de decizie este antrenat pe un subset diferit de date, iar pentru predicțiile tuturor arborilor este calculată o medie pentru a produce predicția finală [96].

În vederea identificării algoritmului care produce cel mai performant model pentru depistarea simptomelor TSA s-a implementat o validare încrucișată (*en.*: Cross-Validation).

Evaluarea statistică a performanței a implicat stocarea valorilor observate reale și a valorilor prezise de modelul de clasificare și a calculării pe baza acestora a metricilor: acuratețea, eroarea de clasificare, coeficientul kappa a lui Cohen și recall. Pentru a determina aceste metrici a fost utilizată matricea de confuzie, în care coloanele reprezintă valorile prezise, iar rândurile reprezintă valorile reale. Conform [97] matricea de confuzie conține următoarele elemente:

- Adevărat Pozitiv (AP): prezicerea corectă a clasei pozitive;
- Adevărat Negativ (AN): prezicerea corectă a clasei negative;
- Fals Pozitiv (FP): prezicerea incorectă a clasei pozitive;
- Fals Negativ (FN): prezicerea incorectă a clasei negative;

Folosind AP, AN, FP și FN au fost determinate metricile:

- **Acuratețea**: Acuratețea reprezintă procentul de predicții corecte din procentul total de predicții și se calculează folosind Ecuația (5.1).

$$A = \frac{AP + AN}{AP + AN + FP + FN} \tag{5.1}$$

- **Eroarea de clasificare**: Eroarea de clasificare reprezintă procentul de predicții incorecte și se calculează folosind Ecuația (5.2);

$$Eroarea\ de\ clasificare = \frac{FP + FN}{AP + AN + FP + FN} \tag{5.2}$$

- **Coeficientul kappa al lui Cohen**: Coeficientul kappa este o metrică statistică robustă, care măsoară acuratețea observată comparativ cu cea așteptată, și se calculează folosind Ecuația (5.5). Unde s-a notat cu (AO) acuratețea observată, iar cu (AA) acuratețea așteptată.

$$AO = \frac{AP + AN}{AP + AN + FP + FN} \tag{5.3}$$

$$AA = \frac{(AP\ +\ FP)(AP\ +\ FN)\ +\ (FN\ +\ AN)(FP\ +\ AN)}{(AP\ +\ FP\ +\ FN\ +\ AN)^2} \tag{5.4}$$

$$Coeficientul\ kappa\ a\ lui\ Cohen = \frac{AO - AA}{1 - AA} \tag{5.5}$$

- **Recall**: Recall sau sensibilitatea tuturor măsurilor este calculată luând media ponderată a recall-ului negativ Ecuația (5.6) și a recall-ului pozitiv Ecuația (5.7) pentru fiecare clasă.

$$Recall\ negativ\ = \frac{AN}{AN + FP} \tag{5.6}$$

$$Recall\ pozitiv\ = \frac{AP}{AP + FN} \tag{5.7}$$

5.4 REZULTATE

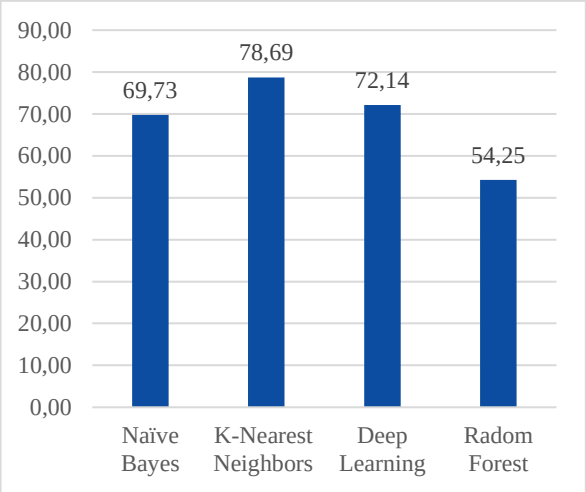


Figura 5.4 Acuratețea produsă de modele antrenate cu algoritmii de învățare automată.

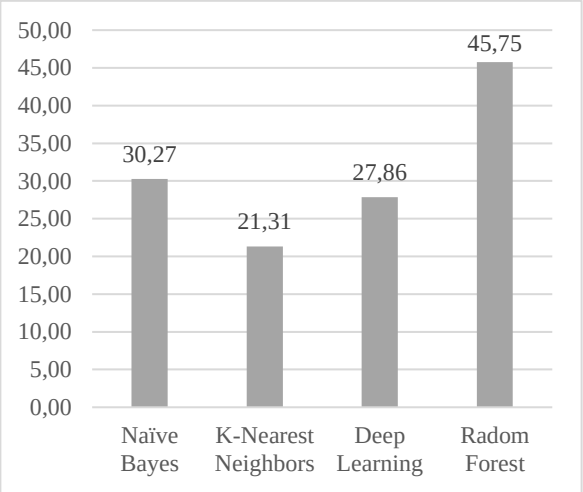


Figura 5.5 Eroarea de clasificare produsă de modele antrenate cu algoritmii de învățare automată.

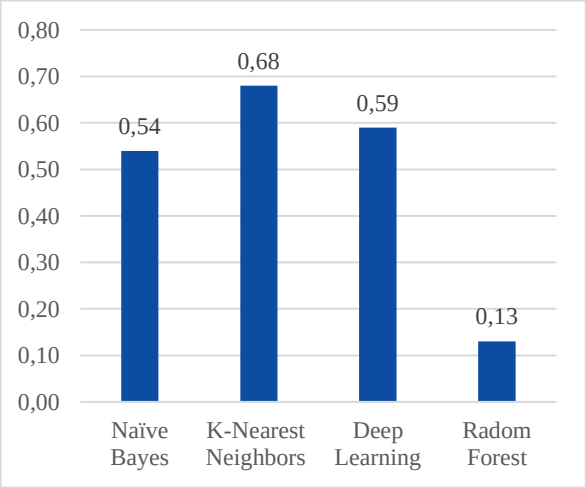


Figura 5.6 Coeficientul kappa al lui Cohen produs de modele antrenate cu algoritmii de învățare automată.

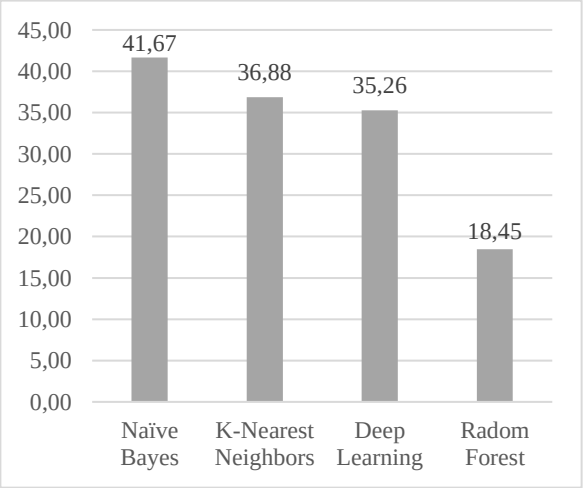


Figura 5.7 Recall produs de modele antrenate cu algoritmii de învățare automată.

Rezultatele obținute în urma determinării acurateței modelului sunt prezentate în Figura 5.4. Algoritmul k-NN a furnizat performanța cea mai ridicată, având valoarea acurateței de 78.69%. S-a constatat că valoarea k, numărul celor mai apropiați vecini, nu influențează semnificativ acuratețea obținută. Pe de altă parte, algoritmul Radom Forest a produs rezultate mai puțin precise, având o acuratețe de doar 54.25%.

Rezultatele obținute în urma determinării erorii de clasificare sunt prezentate în Figura 5.5. Analiza erorii de clasificare a arătat că k-NN este un algoritm eficient pentru identificarea simptomelor autismului producând cea mai mică eroare de 21.31%. Cea mai mare eroare de clasificare de 45.75% a fost determinată de Random Forest.

Rezultatele obținute în urma calculării coeficientul kappa al lui Cohen sunt prezentate în Figura 5.6. Acest coeficient este util în a distinge predicțiile corecte care apar întâmplător, iar o valoare sub 0,40 este puțin satisfăcătoare. k-NN demonstrează cea mai bună legătură, având coeficientul kappa egal cu 0,68.

Rezultatele obținute în urma determinării sensibilității sunt prezentate în Figura 5.7 și contribuie la concluzia că k-NN și Deep Learning sunt algoritmii care produc modele mai performante de identificare a simptomelor TSA.

Procesul de identificare a simptomelor TSA în date text nestructurate, reprezentând răspunsurile părinților la întrebări referitoare la comportamentul copiilor lor diagnosticați cu TSA, a fost unul empiric și a implicat rafinarea iterativă, experimentând cu configurații diferite a caracteristicilor, cu numărul de etichete și cu algoritmii de învățare automată. Rezultatele analizei indicatorilor de performanță au arătat că modelul antrenat folosind algoritmul k-NN produce o acuratețe ridicată de 78.69% și este fezabil pentru identificarea simptomelor TSA. În Exemplu 5.5 este prezentat rezultatul testării modelului antrenat folosind k-NN pe un text furnizat de participantul P_{30} care a fost etichetat cu simptomul „Communication Impairments".

Exemplu 5.5

Question: *Describe your concerns about your child's behavior.*
Response: *The biggest problem was learning to speak again. Up to 1 year and something, the development was normal, then, in 2 months, he stopped talking and just rubbed a small car on the table. I found it very difficult to find a good speech therapist, even though I was doing ABA therapy with an accredited BCBA coordinator. Now he is doing very well. He still does not pronounce the 'r' letter well and we continue with the therapy. I also did 3C therapy and any motor problem we had was solved. He swims and rides a bike!*
Etichetă: Communication Impairments

Rezultatul furnizat de modelul antrenat folosind algoritmul k-NN este:
Predicție: Communication Impairments

Se poate vedea din exemplu, că prin aplicarea text mining s-a identificat corect simptomul TSA „Communication Impairments", în pofida faptului că textul conținea cuvinte și expresii care ar fi putut pune în dificultate modelul antrenat precum „doing very well" sau „motor problem", ce puteau fi asociate cu alte simptome sau cu lipsa indicatorilor autismului. Acest rezultat promițător poate avea implicații semnificative pentru integrarea text mining în aplicații de detecție timpurie a autismului.

Pe măsură ce cercetarea a progresat în explorarea tehnologiei text mining în domeniul medical, s-au identificat câteva implicații importante pentru investigații științifice viitoare:

- Datele trebuie explorate responsabil urmând codul de etică și păstrând confidențialitatea pacienților și familiilor acestora;
- Provocarea minării textelor în alte limbi, decât limba engleză, poate introduce o notă de ambiguitate datorită etapei de traducere. Abordarea acestei limitări necesită o rafinare continuă a algoritmilor pentru a putea furniza ca sursă de antrenare textul în limba originală;
- Diversitatea limitată a setului de date necesită prudență în generalizarea constatărilor. Cercetările viitoare ar trebui să includă o gamă mai variată de expresii lingvistice și contexte culturale pentru a spori aplicabilitatea modelului.

Identificarea automată a simptomelor TSA creează oportunități pentru cercetările viitoare, care pot aprofunda dezvoltarea de instrumente ce folosesc text mining pentru detectarea simptomelor autismului, cu scopul de diagnosticare timpurile și potențial sprijin pentru cadrele medicale și familiile îngrijorate de comportamentul copiilor. Automatizarea procesului de identificare a simptomelor oferă o alternativă eficientă la revizuirea manuală a textelor furnizate de părinți în chestionare, mesaje și video-uri înregistrate în timpul examinărilor medicale.

6 AUTISM AI ADVISOR FOR DIAGNOSIS: SISTEM DE SUPORT PENTRU DIAGNOSTICAREA PRECOCE A TULBURĂRII DE SPECTRU AUTIST

În acest capitol este prezentată metodologia de cercetare și principiile de proiectare și implementare a prototipului aplicației mobile de asistență în diagnostic Autism AI Advisor for Diagnosis. Cunoscută și sub forma acronimului AID, aceasta integrează tehnologia de text mining pentru identificarea simptomelor TSA în date text nestructurate, provenite din răspunsurile părinților la întrebările din testul de screening DISCOVER. În plus, se explorează utilizabilitatea și dezirabilitatea aplicației Autism AI Advisor for Diagnosis, ca și sistem DDSS pentru diagnosticarea TSA, printr-un experiment controlat, la care au participat un număr egal de părinți cu copil cu diagnostic de autism și părinți cu copil fără diagnostic de autism.

6.1 CONTEXT

Practicile curente de diagnosticare TSA sunt consumatoare de timp, testarea clinică poate dura de la 2 până la 3 ore, fapt care reduce semnificativ numărul de pacienți pe care-i poate examina un medic pe zi [98]. În plus, procedurile de diagnosticare a autismului variază foarte mult, iar în consecință timpii de așteptare de la trimitere până la diagnosticare ajung până la 7 luni [99]. Serviciile de e-sănătate, telemedicină și CDSS au potențialul de a îmbunătăți viețile tuturor participanților la actul medical, fie ei furnizori (medici) sau consumatori de sănătate (pacienți) [100]. Tehnologiile pot oferi un acces îmbunătățit la servicii de asistență medicală și pot reduce considerabil timpii de procesare a datelor din formulare și teste clinice.

În urma examinării tehnologiilor utilizate în diagnosticarea TSA (Capitolul 3), s-a observat că tehnologiile digitale și anume aplicațiile mobile, pot constitui cea mai accesibilă formă de instrument medical de sprijin a deciziei [101] din punct de vedere a costurilor de achiziție și a eforturilor de instruire în utilizare. Astfel, cercetarea curentă a propus o abordare nouă în domeniul diagnosticării TSA asistate de tehnologie și introduce Autism AI Advisor for Diagnosis, ce implementează conceptul de text mining ca soluție autonomă de identificare a simptomelor autismului în date text nestructurate, provenite din testele de screening. Soluția software propusă pentru a facilita diagnosticarea TSA are capacități de procesare a limbajului natural și poate identifica simptomele TSA, care sunt manifestate de copii cu vârste fragede cuprinse între 12 luni și 30 de luni, în datele text furnizate de părinții îngrijorați de comportamentul copiilor. Autism AI Advisor for Diagnosis este un tool de suport decizional clinic care are potențialul de a ajuta medicii prin procesarea automată a datelor din testele de screening aferente mai multor pacienți și în același timp oferă părinților informații concrete cu privire la simptomele TSA manifestate de copii.

6.2 METODOLOGIE DE LUCRU

METODOLOGIE DE PROIECTARE

Aplicația Autism AI Advisor for Diagnosis a fost dezvoltată pe baza principiului de proiectare a unui produs software de calitate, care spune că „*pentru a crea o experiență excelentă, trebuie înțeleasă perspectiva utilizatorului*” [102]. Drept urmare, părinții copiilor cu TSA și medicii au fost plasați în centrul procesului de dezvoltare. În locul unei metodologii standard, orientată pe tehnologie folosită în mod obișnuit în dezvoltarea de software, s-a aplicat proiectarea centrată pe om (*en.:* Human-Centered Design (HCD)), în care nevoile oamenilor au avut prioritate față de tehnologie. HCD implică înțelegerea nevoilor utilizatorilor și modul în care proiectarea le poate aborda [103], contribuind astfel la dezvoltare de produse software mai intuitive și eficiente. HCD este recunoscut de cercetătorii din domeniul medical, ca fiind o

metodologie de proiectare eficientă în a obține produse accesibile chiar și pentru persoanele cu tulburări comportamentale precum este TSA [104]. Metodologia de proiectare a urmat pașii HCD din modelul Double Diamond dezvoltat de British Design Council în 2005 [105], care include patru etape (de la E_1 la E_4):

E_1. Etapa de descoperire este de natură divergentă și a implicat cercetarea stadiului actual al tehnologiei în segmentul industriei medicale dedicat diagnosticării TSA, prin explorarea lucrărilor științifice publicate în bazele de date PubMed și ACM Digital Library. Informații importante cu privire la rezultate de ultimă oră în diagnosticarea autismului, dar și limitări însemnate au fost descoperite în urma cercetării.

E_2. Etapa de definire este de natură convergentă, ceea ce înseamnă că este o fază de sintetizare a rezultatelor obținute din cercetarea realizată în E_1. În E_2 au fost definite obiectivele aplicației și au fost identificați utilizatorii finali. Identificarea utilizatorilor este necesară pentru a înțelege pe deplin nevoile și așteptările acestora. Astfel, a fost creată proto-persona pentru doi utilizatori: părinte îngrijorat de comportamentul copilului (Figura 6.1) și medic specializat în diagnosticarea autismului (Figura 6.2).

Termenul „persona” își are rădăcinile în limba latină de la cuvântul „personage”, care a fost utilizat pentru a se referi la o mască purtată de actori în timpul spectacolelor de teatru [106]. În practica de astăzi, proto-personas sunt reprezentări fictive ale utilizatorilor, care sunt aplicate pentru a sprijini echipele de dezvoltare software în activități de creare a cerințelor de experiență a utilizatorului (*en.:* User Experience (UX)). Proto-personas ai utilizatorilor aplicației Autism AI Advisor for Diagnosis au fost create folosind Figma. Figma este un instrument de proiectare, care este special adaptat pentru crearea de interfețe utilizator (*en.:* User Interface (UI)) și proiectarea UX [107]. În procesul de creare s-au analizat caracteristicile demografice ale utilizatorului precum vârsta, genul și ocupația, de asemenea s-au analizat nivelul de cultură digitală, competențele în utilizarea tehnologiilor moderne și obiectivele acestora.

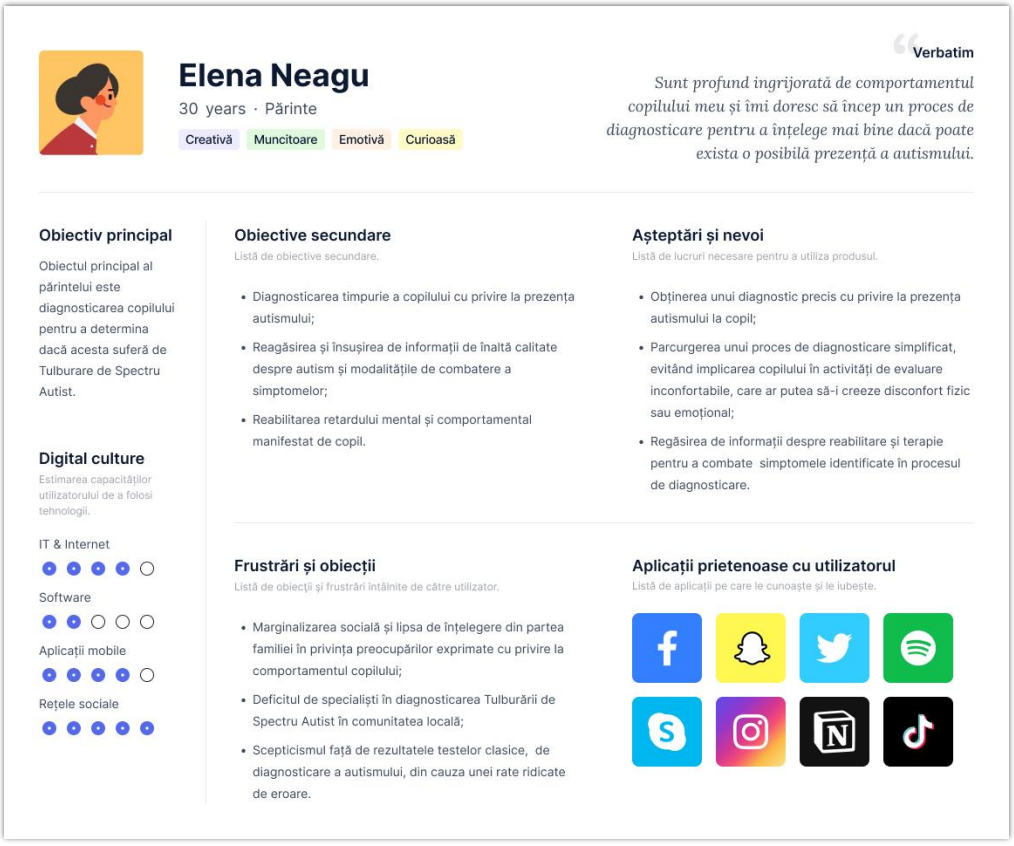


Figura 6.1 Proto-persona părinte creată folosind Figma.

Figura 6.2 Proto-persona medic creată folosind Figma.

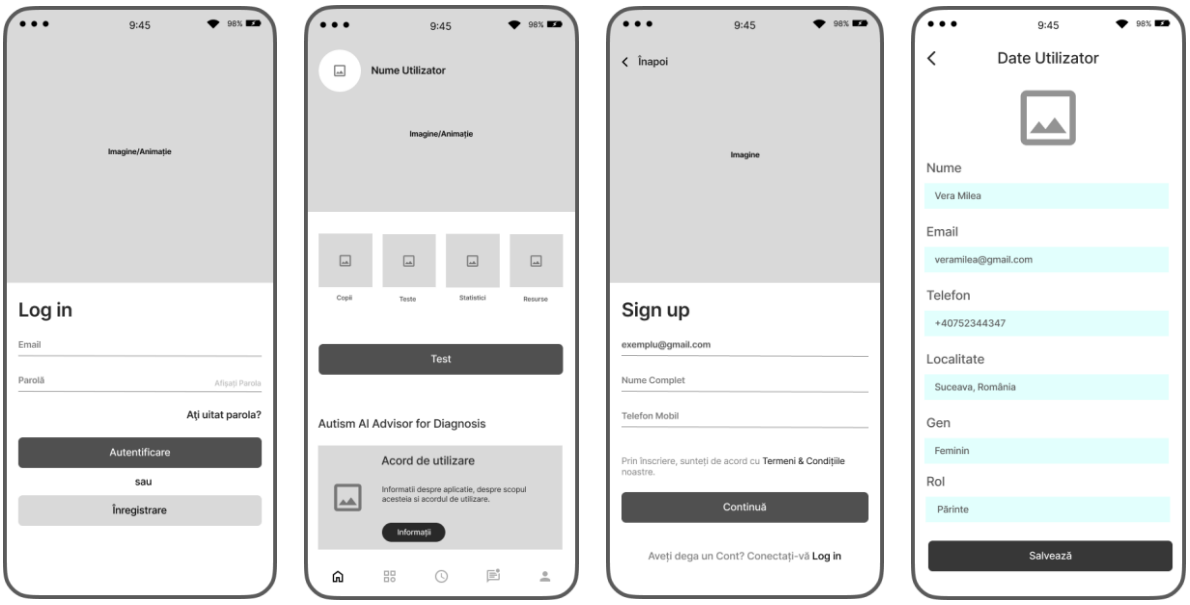
Figura 6.1 prezintă proto-persona unui părinte îngrijorat de comportamentul copilului său. Părintele își dorește să parcurgă un proces de diagnosticare simplificat, care să nu expună copilul la activități de evaluare inconfortabile efectuate în condiții clinice, care ar putea să-i declanșeze comportamente cu valoare de criză.

Figura 6.2 ilustrează proto-persona unui medic specializat în diagnosticarea TSA, a cărui obiectiv principal este identificarea corectă a simptomelor autismului pentru stabilirea unui diagnostic precis. Medicul are abilități dezvoltate în utilizarea instrumentelor digitale, însă se confruntă cu lipsa resurselor tehnologice.

Efortul de a înțelege utilizatorii finali a facilitat identificarea caracteristicilor esențiale pentru o aplicație de screening eficientă precum: interfața UI intuitivă, accesibilitatea informațiilor prezentate și securitatea datelor.

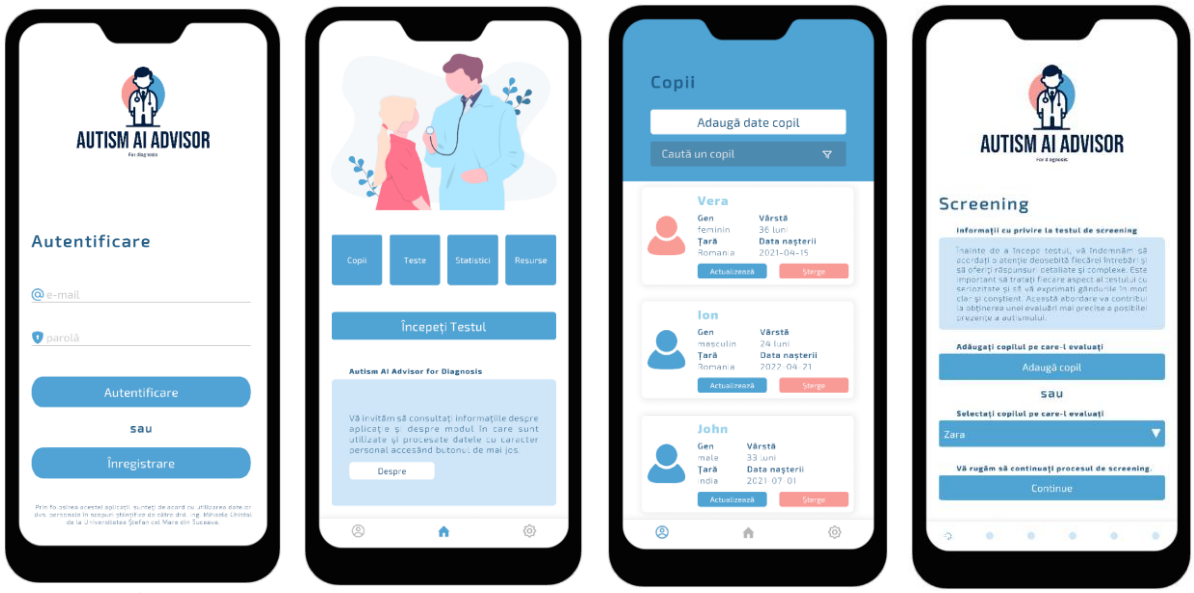
E3. Etapa de dezvoltare este de natură divergentă și a implicat conceptualizarea aplicației și iterarea ideilor, care sunt fundamentale pentru procesul de creare și rafinare a soluției tehnice. În această etapă, au fost create wireframe-urile aplicației folosind Figma, iar câteva dintre acestea sunt prezentate în Figura 6.3. Wireframe-ul reprezintă un instrument puternic care traduce cu acuratețe ideile într-o reprezentare vizuală, ce exclude în mod deliberat culoarea, grafica și fonturile stilizate. Wireframe-urile facilitează de asemenea identificarea acțiunilor pe care le va întreprinde utilizatorul, permițând ierarhizarea elementelor vizuale în mod optim.

E4. Etapa de livrare este de natură convergentă și a presupus transformarea conceptului identificat prin proces de cercetare și iterare, într-un produs tangibil. Aceasta a implicat definirea caracteristicilor aplicației, identificarea instrumentelor de dezvoltare, proiectarea arhitecturii și implementarea tehnică. Figura 6.4 prezintă capturi de ecran din aplicația implementată pe baza ideilor din conceptul final.



a) b) c) d)

Figura 6.3 Wireframe-urile aplicației Autism AI Advisor For Diagnosis a) Ecran de autentificare; b) Ecran principal; c) Ecran de înregistrare; și d) Date utilizator.



a) b) c) d)

Figura 6.4 Capturi de ecran din Aplicația Autism AI Advisor For Diagnosis a) Ecran de autentificare; b) Ecran principal; c) Ecran de management copii; și d) Ecran de screening.

ARHITECTURA SOFTWARE

Autism AI Advisor For Diagnosis este o aplicație mobilă, concepută pentru a facilita procesul de diagnosticare a TSA la copii mici, cu vârstă cuprinsă între 12 luni și 30 de luni. Dezvoltată cu o abordare centrată pe om, AID integrează algoritmi avansați de inteligență artificială pentru a analiza datele text nestructurate provenite din răspunsurile părinților la întrebările din testele de screening, cu scopul de a determina probabilitatea ca copilul evaluat să sufere de autism. Obiectivul principal al Autism AI Advisor for Diagnosis este de a împuter-

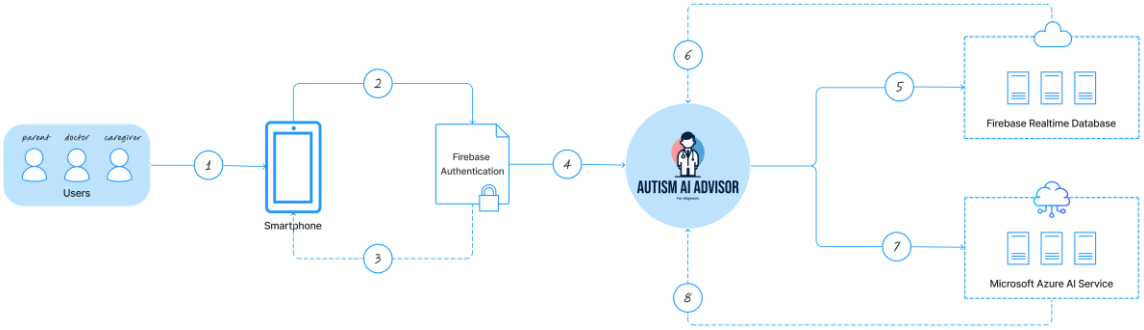


Figura 6.5 Schema bloc a aplicației Autism AI Advisor For Diagnosis cu prezentarea componentelor software și fluxurilor de date.

nici părinții și profesioniștii din domeniul sănătății cu un instrument eficient în identificarea timpurie a simptomelor TSA, facilitând astfel intervenția terapeutică.

Figura 6.5 ilustrează arhitectura aplicației Autism AI Advisor For Diagnosis, care a fost creată folosind Figma. Schema bloc oferă o imagine de ansamblu la nivel înalt a componentelor software. Sistemul cuprinde patru componente de bază: (1) Aplicația mobilă, dedicată pentru dispozitive și smartphone-uri cu sistem de operare Android, dezvoltată folosind motorul multiplatformă Unity; (2) Baza de date NoSQL implementată folosind Firebase Realtime Database care este găzduită în cloud; (3) Serviciul de autentificare dezvoltat prin utilizarea Firebase Authentication; (4) Servicii de inteligență artificială cu capacități de text mining și stocate în cloud, pentru care s-a folosit Microsoft Azure AI Service. Arhitectura include și utilizatori finali: părinte, îngrijitor și medic.

IMPLEMENTARE TEHNICĂ

Autism AI Advisor For Diagnosis este o aplicație de asistență în diagnostic care oferă medicilor suport informat despre simptomele TSA identificate la copilul examinat, prin analiza automată a datelor din testele de screening. Figura 6.6 prezintă ecranul principal al aplicației AID, care oferă utilizatorului posibilitatea de a adăuga datele mai multor copii, de a evalua prin teste de screening starea de sănătate a copiilor și de a vizualiza rezultatele testelor.

Implementarea aplicației Autism AI Advisor for Diagnosis a implicat mai întâi dezvoltarea unui nou test de screening, numit Depistarea Inteligentă a Simptomelor Comportamentale Observate în Evaluarea pentru Riscul Autismului (DISCOVER).

Tabel 6.1 Testul de screening DISCOVER.

Număr	Întrebare
Q_1	Vă rugăm să descrieți comportamentul copilului dumneavoastră care vă ridică îngrijorări, inclusiv detaliile specifice ale acestui comportament și problemele pe care le-ați identificat în legătură cu acesta.
Q_2	Vă rugăm să descrieți modul în care copilul interacționează social și comunică.
Q_3	Ați observat comportamente repetitive sau stereotipe la copilul dumneavoastră? Vă rugăm să povestiți despre aceste comportamente.
Q_4	Vă rugăm să descrieți modul în care copilul dumneavoastră reacționează la stimuli senzoriali, cum ar fi sunetele, texturile sau lumina.
Q_5	Cum comunică copilul dumneavoastră atât verbal, prin formularea propozițiilor, cât și nonverbal, prin gesturi sau expresii faciale? Vă rugăm să oferiți exemple specifice care să ilustreze modul în care copilul comunică.

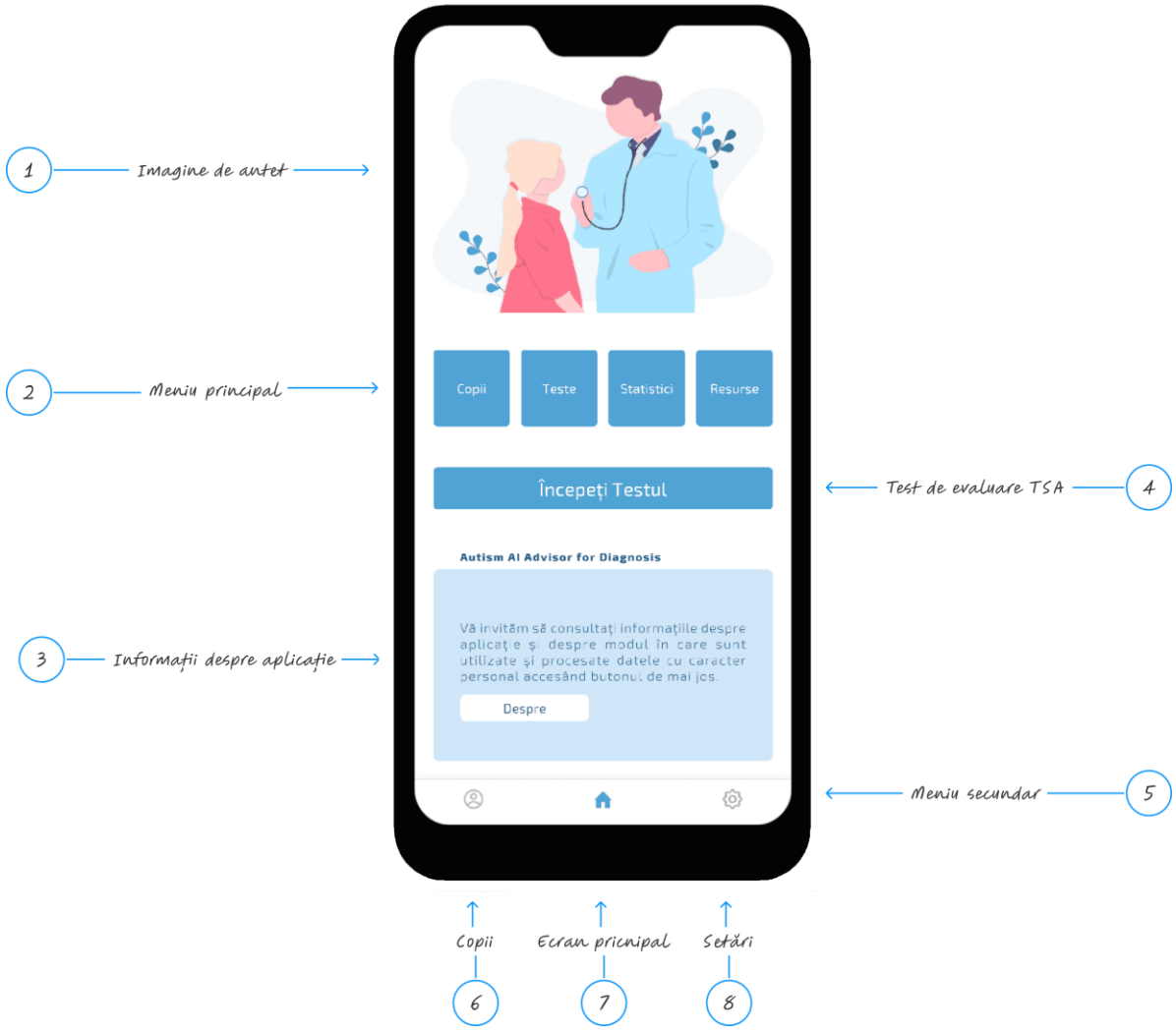


Figura 6.6 Ecranul principal al aplicației Autism AI Advisor for Diagnosis.

Testul este prezentat în Tabelul 6.1 și conține 5 întrebări cu răspuns liber. Întrebările au fost redactate pe baza recomandărilor medicale de diagnosticare a TSA descrise în DSM-5.

Autism AI Advisor for Diagnosis integrează și testul standardizat Quantitative CHecklist for Autism in Toddlers (Q-CHAT), pus la dispoziție în limba romană de către Autism Research Centre de la Universitatea din Cambridge [108]. Q-CHAT conține 25 de întrebări cu răspuns multiplu și oferă un grad ridicat de încredere.

BAZA DE DATE

Gestionarea datelor de screening a fost obținută prin implementarea unei baze de date folosind Firebase Realtime Database. Baza de date este stocată pe cloud și este de tip NoSQL, caracteristici care o fac mai optimizată comparativ cu o bază de date relațională. În loc de solicitările HTTP tipice, Firebase Realtime Database utilizează sincronizarea datelor, de fiecare dată când acestea se modifică orice dispozitiv conectat primește această actualizare în milisecunde [109].

APLICAȚIA MOBILĂ

Dezvoltarea aplicației mobile Autism AI Advisor For Diagnosis a fost realizată prin utilizarea motorului de joc Unity, versiunea de editor 2022.3.21f1, și Microsoft Visual Studio

Community 2022, versiunea 17.9.5. Motorul Unity are capacități excepționale pentru dezvoltare de aplicații 2D și pentru migrarea acestora pe diferite sisteme de operare precum Android și iOS.

Arhitectura principală a aplicației este Model-View-Presenter (MVP) și a fost aleasă pentru că este un model arhitectural de UI care facilitează creșterea capacității de întreținere a codului și implementarea testării automate [72]. Modelul (*en.*: model), este o interfață care definește datele ce vor fi afișate sau manipulate de către utilizator. Vizualizarea (*en.*: view) este o interfață care afișează datele în elementele de UI. Prezentatorul (*en.*: presenter) joacă rolul unui intermediar care preia comenzile din vizualizare și acționează asupra modelului.

Zenject, versiunea 9.2.0, a fost integrată pentru a asigura un cadru ușor de injectare a dependențelor. Zenject este un framework de injecția de dependențe creat special pentru Unity.

SERVICIU DE AUTENTIFICARE

Confidențialitatea și securitatea datelor utilizatorilor au fost obținute prin implementarea posibilității de înregistrare și autentificare cu e-mail și parolă. Pentru aceasta s-a utilizat SDK-ul Firebase Authentication care oferă metode pentru a crea și gestiona utilizatori. Datele utilizatorilor sunt stocate în cloud, ceea ce garantează aceiași experiență pe toate dispozitivele.

SERVICII CLOUD DE AI

Aplicația Autism AI Advisor For Diagnosis integrează două servicii Microsoft Azure AI pentru analiză avansată a textului și capabilități de procesare a limbajului natural. Microsoft Azure AI Language este un serviciu gestionat în cloud care oferă funcții de explorare și analiză a textului [110]. Azure AI Language Service a fost utilizat pentru a identifica simptome TSA prin interpretarea răspunsurilor părinților la întrebările testului DISCOVER. Proiectarea și antrenarea modelului a fost realizată folosind Microsoft Azure Language Studio care are capacități avansate de text mining. Pentru stocarea în siguranță a datelor, utilizate de serviciile AI, a fost integrat Microsoft Azure Storage care este o soluție de stocare în cloud.

MODELULUI DE AI

Proiectarea și implementarea procesului autonom de identificare a simptomelor TSA în date text nestructurate (Figura 6.7) a implicat dezvoltarea unui model AI. Dat fiind specificul sarcinii propuse, s-a optat pentru antrenarea unui model de clasificare folosind învățarea supervizată total. Modelul AI a fost antrenat cu ajutorul capacităților computaționale ale Microsoft Azure Language Studio și a constat în parcurgerea a cinci etape:

- **E_1** Crearea setului de date;
- **E_2** Etichetarea setului de date;
- **E_3** Antrenarea modelului;
- **E_4** Evaluarea performanței;
- **E_5** Implementarea modelului.

E_1. Etapa de creare a setului de date a constat în utilizarea datelor colectate de la N=44 de participanți la experimentul descris în Capitolul 5. Setul de date conține 473 de înregistrări care au fost procesate manual și curățate de zgomot. Spre deosebire de abordarea anterioară, datele nu au fost traduse din limba română în limba engleză, deoarece acuratețea modelului poate fi afectată. Stocarea datelor a implicat salvarea individuală a fiecărei înregistrări în câte un document text cu extensia .txt și apoi încărcarea acestuia în Microsoft Azure Storage.

E_2. Etapa de etichetare a setului de date este un factor cheie în determinarea performanței modelului. Clasele trebuie să fie clar distinctibile și separabile pentru a evita ambiguitatea.

Limitările identificate în cercetarea prezentată în Capitolul 5 au demonstrat că setul de date este mai degrabă potrivit pentru un proces de clasificare cu mai multe etichete.

Figura 6.7 Procesul autonom de identificare a simptomelor TSA folosind aplicația Autism AI Advisor For Diagnosis.

Schema de etichetare conține 19 etichete, dintre care 18 etichete reprezintă simptome specifice autismului și o etichetă specială „Asimptomatic" care semnifică lipsa simptomelor. Documentele din setul de date au fost etichetate în mai multe clase folosind interfața UI pentru etichetare a Microsoft Azure Language Studio. Distribuția etichetelor nu este uniformă, cele mai comune apariții le au etichetele asimptomatic 38.90%, dificultăți de comunicare 25.75% și deficiențe în reciprocitatea socio-emoțională 11.46%. Incidența ridicată a acestor etichete este explicabilă deoarece simptomele aferente sunt considerate manifestări principale ale TSA [111]. Schema de etichetare a fost salvată într-un document cu formatul JSON.

E_3. Etapa de antrenare a modelului a implicat utilizarea setului de documente clasificate și împărțirea acestuia în documente de antrenare 80% și documente de testare 20%. Modelul personalizat de clasificare a textului a fost învățat să distingă simptomele TSA prin aplicarea algoritmilor de învățare automată într-un job de formare creat în Microsoft Azure Language Studio. Timpul mediu de instruire a fost la nivelul zecilor de minute.

E_4. Evaluarea performanței pentru modelul de clasificare a fost realizată imediat după antrenarea acestuia și a constat în prezicerea simptomelor TSA pentru documentele setului de testare. Folosind etichetele prezise și comparându-le cu clasele corecte, s-au determinat următoarele metrici de performanță: precizia, recall și scorul F1.

Precizia (*en.*: precision (p)) modelului este de 91.57% și a fost calculată folosind Ecuația (6.1) [112]. Valoarea semnificativă a acesteia indică că modelul este exact și are o performanță ridicată în evitarea prezicerilor fals pozitive ale simptomele TSA.

$$\mathrm{p} = \frac{AP}{AP + FP} \tag{6.1}$$

Recall (*en.*: recall (r)) indică o valoare de 84.44% și a fost calculată folosind Ecuația (6.2) [112]. Valoarea demonstrează performanța modelului în evitarea prezicerilor fals negative.

$$\mathrm{r} = \frac{AP}{AP + FN} \tag{6.2}$$

Scorul F1 este de 87.86% și este o metrică care reprezintă o măsură generală a performanței modelului. Aceasta este media armonică a preciziei și recall-ului și a fost calculată folosind Ecuația (6.3) [112] reprezentată mai jos:

$$\mathrm{F1} = \frac{2 * p * r}{p + r} \tag{6.3}$$

E_5 Etapa de implementare a modelului a constat în lansarea și găzduirea acestuia în regiunea de est a Statelor Unite ale Americii (SUA).

Integrarea în aplicația Autism AI Advisor for Diagnosis a fost realizată prin importarea Azure SDK, pentru .NET, în Unity și interogarea modelului prin API-ul de predicție. Conectarea la serviciu a fost realizată prin crearea unui TextAnalyticsClient pentru care s-au specificat endpoint-ul și credențialele. Clasificarea răspunsurilor părinților la întrebările testului DISCOVER a fost realizată prin apelul funcției asincrone MultiLabelClassifyAsync, care returnează un obiect de tipul ClassifyDocumentOperation și care conține informații relevante cu privire la clase și scorul de încredere a prezicerii. Rezultatele identificării simptomelor au fost stocate în baza de date Firebase Realtime Database și au fost puse la dispoziția utilizatorilor prin interfața numită „Teste”. Prin intermediul acestei interfețe, părinții și cadrele medicale pot vizualiza testele efectuate pentru fiecare copil adăugat în contul de utilizator.

6.3 EXPERIMENT

Este important ca diagnosticarea autismului să fie realizată cât mai timpuriu pentru a obține rezultate pozitive în procesul de recuperare terapeutică. Membrii familiei nucleare sunt primii care observă întârzierile în dezvoltarea cognitivă și comportamentele atipice ale copiilor. Astfel, implicarea în experiment a adulților care au cel puțin un copil este justificată.

Participanți: În mod voluntar la experiment au ales să participe N=22 de participanți. Participanții au fost distribuiți în două grupuri reprezentative pentru obiectivul propus, și anume, grupul „*părinți cu copil cu diagnostic de autism*” (G_1) și grupul „*părinți cu copil fără diagnostic de autism*” (G_2). Distribuția participanților este uniformă, cu o divizare egală de 11 participanți (50%) aparținând G_1 și 11 participanți (50%) aparținând G_2.

Datele demografice sugerează că majoritatea participanților de 15 (71.4%) locuiesc în mediul urban în timp ce restul 6 (28.6%) în mediul rural. Statutul socio-profesional raportat de participanți este divers: 11 (52.4%) sunt angajați cu normă întreagă, 1 (4.8%) angajat cu o fracțiune de normă, 5 (23.8%) persoane casnice, 2 (9.5%) liberi profesioniști și 2 (9.5%) asistenți personali ai copilului lor.

Cercetările contemporane demonstrează că părinții care au avut un copil cu TSA au fost mai susceptibili să identifice simptomele autismului la următorii copii [66]. Această distincție a determinat explorarea datelor demografice ale copiilor. Majoritatea părinților 11 (52.4%) au doi copii, urmați de 5 (23.8%) cu un copil și 3 (14.3%) cu trei copii, restul de 2 (9.5%) părinți având peste 3 copii. Doar un părinte a raportat că are mai mult de 1 copil diagnosticat cu TSA. Autismul nu este distribuit în mod egal între genuri, cu un raport între masculin și feminin de 4:1 [113]. În studiul nostru raportul dintre genuri nu este egal fiind 15:7, 68.2% din copiii participanților au genul masculin și 31.8% au genul feminin, ceea ce se aliniază cu statisticile de incidență raportate la nivel global.

Sarcină: Experimentul a fost alcătuit din trei faze:

1. **Faza de configurare**: participanții au fost rugați să-și instaleze aplicația Autism AI Advisor For Diagnosis pe telefon folosind Google Play. Dacă participantul nu dispunea de un device cu sistem de operare Android, atunci i se punea la dispoziție telefonul Motorola Moto G84 5G, modelul XT2347-2, care avea aplicația instalată;

2. **Faza de testare**: participanții au interacționat cu aplicația și au răspuns la testele de screening DISCOVER și Q-CHAT. În Figura 6.8 sunt prezentate câteva fotografii realizate în timpul experimentului, care demonstrează modul în care participanții au interacționat cu aplicația Autism AI Advisor for Diagnosis;
3. **Faza de evaluare**: participanții au completat chestionarul pentru evaluarea aplicației Autism AI Advisor for Diagnosis folosind versiunea web a Google Forms.

Instrument: Chestionarul a fost dezvoltat pentru a evalua trei domenii-cheie: variabile demografice și psihologice ale părintelui, variabile care descriu percepția aplicației și variabile legate de utilizabilitatea diagnosticării TSA asistată de aplicația Autism AI Advisor For Diagnosis. Chestionarul a cuprins următoarele măsuri menite să capteze informații specifice fiecărui domeniu:

1. **Măsuri ale experienței părintelui***:*
   (a) Computer Self-Efficacy Scale (CSE) [114] este o scală de auto-eficacitate în utilizarea tehnologiilor moderne, precum calculatorului, care a permis determinarea variabilei:
      (i) nivel de încredere în utilizarea tehnologiilor.
   (b) Autism Parenting Stress Index (APSI) [115] este un instrument care evaluează nivelul de stres perceput de părinți în legătură cu îngrijirea unui copil cu autism. APSI analizează aspecte legate de comportamentul copilului, relațiile sociale și aspecte fiziologice, și determină variabila:
      (i) nivel de stres.
2. **Măsuri ale percepției aplicației**:
   (a) Chestionarul de Percepție a Aplicației este inspirat din cercetările lui V. Venkatesh [116] și P. Zhang et al. [117], și conține 16 afirmații pentru evaluarea cărora s-au folosit scale Likert în 7 puncte. Chestionarul determină cinci variabile importante raportate la aplicația Autism AI Advisor For Diagnosis:
      (i) satisfacția/plăcerea;
      (ii) utilitatea;
      (iii)ușurința utilizării;
      (iv)percepția controlului;
      (v) intenția comportamentală de a folosi aplicația.
3. **Măsuri ale utilizabilității aplicației**:
   (a) System Usability Scale (SUS) [118] este o scală de evaluare a gradului de utilizabilitate al unui sistem. Acest chestionar standardizat a fost utilizat pentru a evalua percepțiile subiective ale utilizatorilor cu privire la capacitatea de utilizare a aplicației Autism AI Advisor For Diagnosis, oferind feedback valoros asupra eficienței și satisfacției. Variabila care este determinată utilizând SUS este:
      (i) nivel de utilizabilitate.
   (b) Chestionar de Fezabilitate a Caracteristicilor Aplicației, ce conține 10 afirmații pentru evaluarea cărora s-au folosit scale Likert în 5 puncte. Chestionarul determină variabile raportate la funcționalitățile aplicației Autism AI Advisor For Diagnosis precum:
      (i) utilitatea UI/UX;
      (ii) utilitatea opțiunilor de înregistrare;
      (iii)utilitatea opțiunilor de autentificare;
      (iv)utilitatea opțiunilor de management copii;
      (v) utilitatea opțiunilor de screening;
      (vi)fezabilitatea sistemelor de semnalare erori, de progres și notificare.

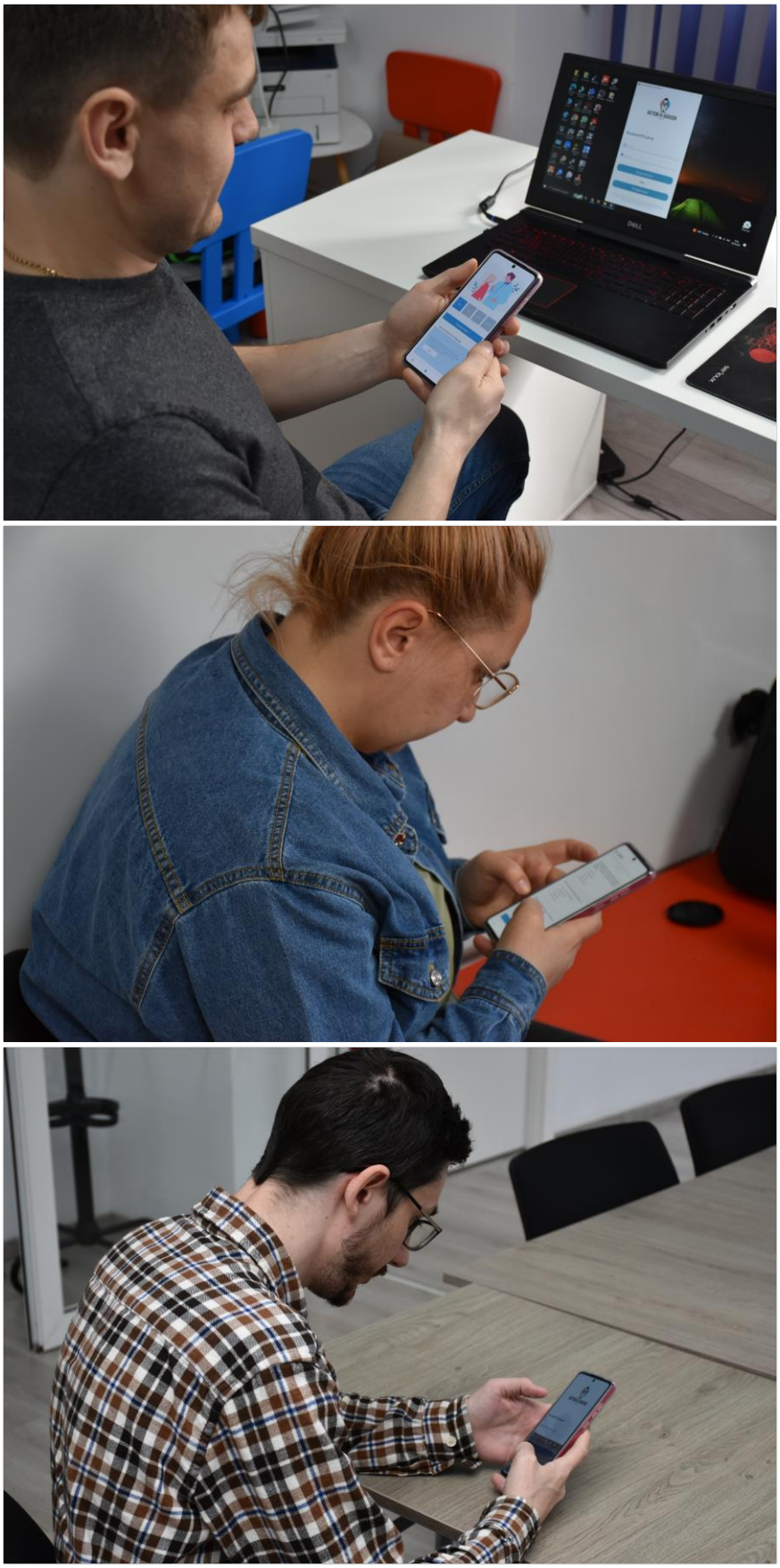


Figura 6.8 Fotografii realizate în timpul experimentului care prezintă trei participanți interacționând cu aplicația Autism AI Advisor for Diagnosis.

6.4 REZULTATE

REZULTATE UTILIZABILITATE

Evaluarea empirică a utilizabilității prototipului de tehnologie Autism AI Advisor for Diagnosis, propus pentru diagnosticarea TSA, a dezvăluit perspective importante.

Experiența părintelui: Pentru a înțelege abilitățile participanților în utilizarea tehnologiilor moderne a fost administrată scala de auto-eficacitate CSE. Douăzeci de participanți (90.90%) au obținut un nivel de încredere în utilizarea tehnologiilor „*foarte încrezător*"; doi (9.09%) „*moderat încrezător*"; niciun participant nu a avut un nivel de încredere „*deloc încrezător*".

Analiza indicelui de stres parental a fost realizată comparativ pentru a distinge între percepțiile grupurilor G_1 și G_2, și s-a utilizat instrumentul APSI care conține 13 afirmații pentru evaluarea cărora s-au folosit scale Likert în 5 puncte de la 0 (*deloc stresant*) la 4 (*atât de stresant încât simt că nu mai găsesc soluții*). Scorul total variază de la 0 la 52; scorurile mai mari indicând un stres parental mai mare. Participanții din G_1 au raportat un nivel de stres parental mai mare, cu o medie de 9,36 comparativ cu participanții din G_2 a căror medie este de 8,27. Totuși, rezultatele cu privire la stresul perceput de părinți în legătură cu îngrijirea unui copil cu autism nu indică valori asociate cu stres puternic.

Percepția aplicației: Percepția aplicației a fost evaluată printr-un chestionar cu 16 itemi pentru măsurarea cărora s-au folosit scale Likert în 7 puncte de la 0 (*dezacord puternic*) la 6 (*acord puternic*).

Satisfacția/plăcerea utilizării aplicației este una pozitivă, participanții raportând un nivel înalt de mulțumire după cum se poate observa în Figura 6.9 (a).

Utilitatea aplicației în procesul de screening și diagnosticare TSA a fost percepută ca fiind bună (Figura 6.9 (b)). Participații au raportat că aceasta poate fi un instrument de sprijin în evaluarea simptomelor autismului care să contribuie și la conștientizarea comportamentelor atipice.

Ușurința utilizării este percepută pozitiv cu privire la facilitatea cu care este utilizată aplicația, intuitivitatea și efortul mental necesar pentru a o folosi (Figura 6.9 (c)).

Percepția controlului este descrisă ca fiind bună, majoritatea participanților au simțit că au control asupra procesului de utilizare a aplicației și a modului în care aceasta își îndeplinește funcțiile (Figura 6.9 (d)).

Intenția comportamentală de a folosi aplicația indică o predispunere pozitivă a participanților pentru a continua utilizarea aplicației în viitor și intenția de a recomanda instrumentul de screening și altor persoane (Figura 6.9 (e)).

Utilizabilitatea aplicației: Pentru a înțelege utilizabilitatea aplicației Autism AI Advisor For Diagnosis a fost folosită scala SUS. Analiza rezultatelor SUS au dezvăluit un nivel bun de utilizabilitate percepută (SUS=81.36%). Scorul SUS este peste medie, aproape de pragul „excelent" de 85 sugerat de Bangor et al. [119].

Fezabilitatea caracteristicilor și funcționalităților aplicației a fost evaluată printr-un chestionar ce conține 10 afirmații, pentru evaluarea cărora s-au folosit scale Likert în 5 puncte de la 0 (*foarte mică*) la 4 *(foarte mare*). Analiza răspunsurilor participanților a dezvăluit un rezultat surprinzător, toate variabilele evaluate (*UI/UX, opțiuni de înregistrare, autentificare, management copii și screening, sistemele de semnalare erori, progres și notificare*) au fost apreciate ca fiind de o importanță și utilitate „foarte mare". Această evaluare, subliniază percepția pozitivă a utilizatorilor cu privire la aplicația Autism AI Advisor For Diagnosis.

Studiul curent a cercetat și comportamentul participanților în utilizarea tehnologiilor de diagnosticare TSA. Majoritatea covârșitoare de 21 (95.5%) dintre participanți au considerat importantă diagnosticarea timpurie a TSA; doar 1 (4.5%) participant a evaluat-o ca fiind neimportantă. Screening-ul autismului asistat de tehnologii a fost perceput de participanți ca

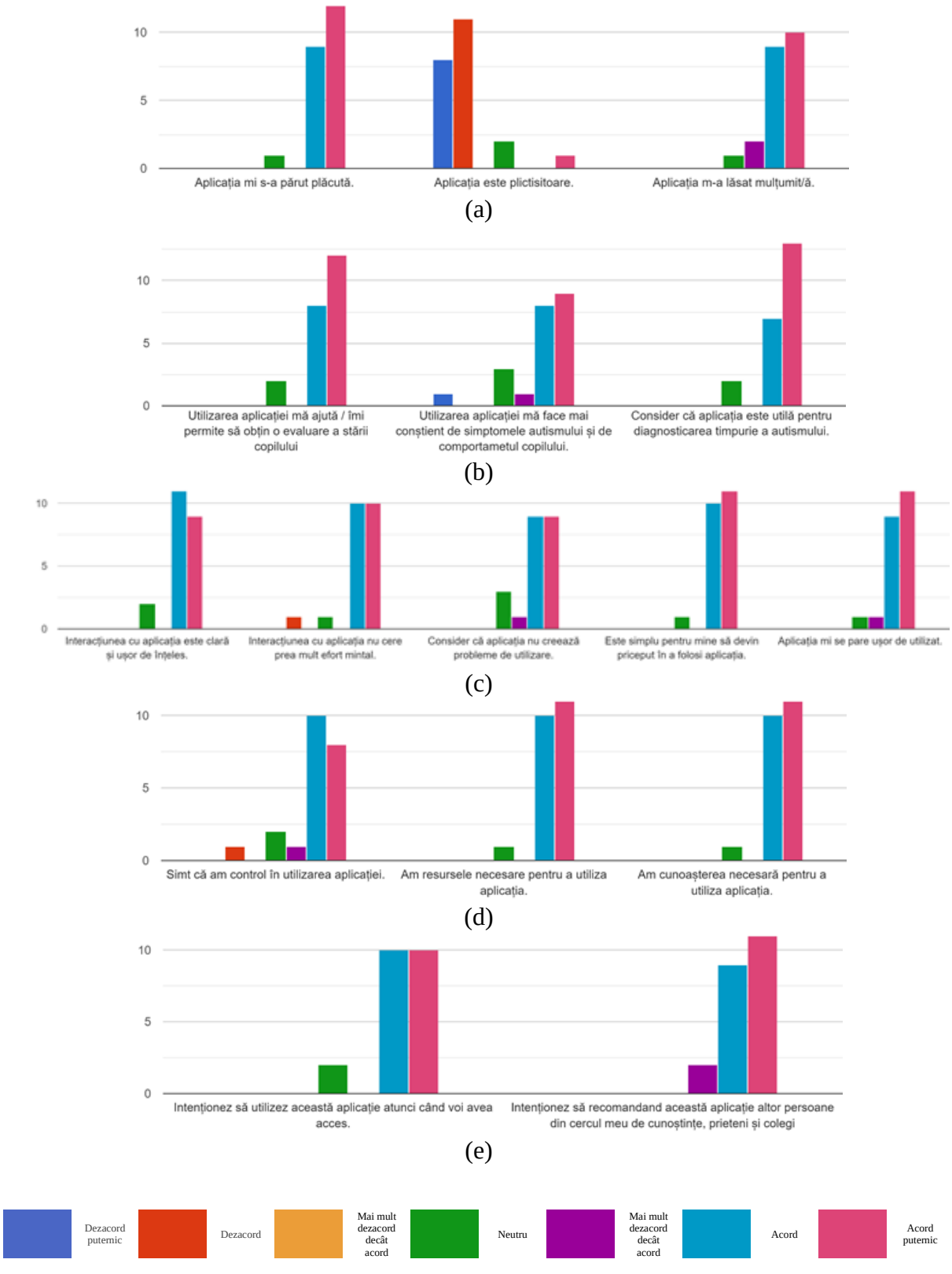


Figura 6.9 Percepția participanților despre aplicația Autism AI Advisor for Diagnosis prezentând variabilele a) Satisfacția/plăcerea; b) Utilitatea; c) Ușurința utilizării; d) Percepția controlului; e) Intenția comportamentală de a folosi aplicația.

fiind util 16 (72.7%). Nivelul de confort în utilizarea tehnologiilor moderne pentru diagnosticarea TSA a fost raportat ca fiind sporit 16 (72.7%). În general, participanții au demonstrat o atitudine pozitivă față de diagnosticarea TSA asistată de tehnologie, deoarece consideră că aceasta poate minimiza expunerea la prejudecățile membrilor societății, asociate cu diagnosticul de autism.

REZULTATE PERFORMANȚĂ

Răspunsurile participanților din grupurile G_1 și G_2 la întrebările testului DISCOVER au fost analizate folosind metode de explorare a textului pentru a identifica simptome specifice autismului. Performanța calculată a modelului AI este ridicată, indicând o precizie de 91.57% și este expusă împreună cu alte metrici în Capitolul 6.2. În plus, pentru a înțelege eficiența în condiții reale a aplicației, studiul a analizat comparativ rezultatele obținute pentru cele două teste: DISCOVER și Q-CHAT. Metoda de screening DISCOVER, bazată pe text mining, a identificat corect starea de sănătate în ceea ce privește TSA la 68.18% (15/22) dintre copii evaluați, iar metoda Q-CHAT a identificat corect la 68.18% (15/22). Identificarea corectă înseamnă că diagnosticul TSA oferit de metoda de screening este același cu starea de sănătate a copilului raportată de părinte și validată de către medic. Rezultatele egale indică că modelul AI bazat pe text mining poate fi folosit independent în practica medicală pentru a prezice TSA deoarece produce efecte similare.

Explorarea performanței a abordat și analiza matricei de confuzie aferentă testului DISCOVER și Q-CHAT prezentată în Figura 6.10. DISCOVER a produs o eficiență mai bună cu privire la posibilitatea ca un copil să sufere de TSA pentru participanții din G_1, având valoarea pentru adevărat pozitiv egală cu 9. La polul opus Q-CHAT a identificat mai bine absența TSA la participanții din G_2, având valoarea pentru adevărat negativ egală cu 9. Analiza rezultatelor fals negative a demonstrat o performanță mai bună produsă de DISCOVER (FN=3) comparativ cu Q-CHAT (FN=5). Rezultatele fals negative pot avea o consecință gravă pentru o aplicație medicală și pot contribui la subdiagnosticarea pacienților. În ceea ce privește rezultatele fals pozitive Q-CHAT a produs o valoare FP=2, în timp ce DISCOVER a identificat FP=4. Aceste rezultate sugerează necesitatea validării diagnosticului TSA de către un specialist medical.

Performanța testelor de screening este puternic corelată cu percepția părintelui despre comportamentul copilului. În cazul testului DISCOVER un impact îl are și abilitatea părintelui de a descrie manifestările copilului printr-un text.

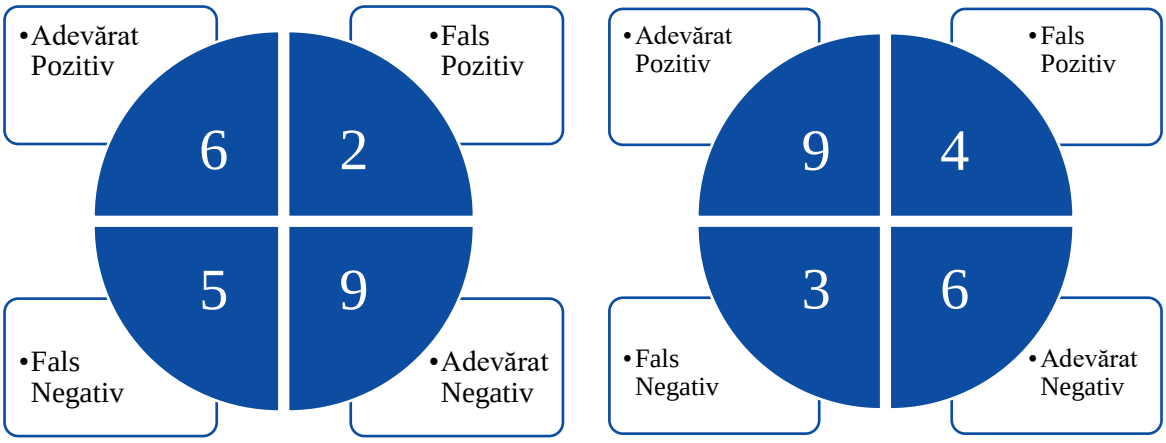


Figura 6.10 Matricea de confuzie pentru testul de screening: a) Q-CHAT; b) DISCOVER.

Figura 6.11 prezintă răspunsul participantului P_2 la întrebarea Q_2. Răspunsul participantului P_2 este detaliat și oferă informații generoase despre abilitățile de socializare a copilului ceea ce a favorizat clasificarea corectă a acestuia în clasa „asimptomatic”.

> „*Darius este în general foarte atent la persoanele din jur. Este destul de timid când întâlnește persoane străine, dar cu cei cunoscuți este foarte interesat în a avea un dialog. În general este foarte implicat atunci când o persoană din familie sau cunoscută se joacă sau comunică cu el. Este în general vesel și jucăuș, privește persoanele în ochi, zâmbește, râde, scoate sunete, imită și face diverse grimase în funcție de specificul dialogului. Aș spune că este un copil foarte comunicativ.”*

Figura 6.11 Răspunsul participantului P_4 la întrebarea Q_2 din testul DISCOVER.

Figura 6.12 prezintă răspunsul participantului P_{10} la întrebarea Q_3. Acesta a fost clasificat greșit în clasa „asimptomatic” deși ar fi trebuit asociat cu simptomul „Echolalia”, care este o caracteristică tipică TSA prezentă la 75%-80% dintre indivizii verbali și care se manifestă prin repetiția fără sens a cuvintelor [120]. Rezultatul incorect nu este asociat cu lungimea răspunsului, exprimată în număr de caractere, deoarece în medie răspunsurile participanților clasificate corect au avut un număr de caractere de 125,98, iar răspunsurile clasificate greșit de 152,33. Analiza setului de date a dezvăluit că aceasta poate fi o limitare asociată cu un număr mic de documente etichetate cu clasa „Echolalia” (1.90%). Astfel, acuratețea modelului AI ar putea fi optimizată prin îmbogățirea setului de antrenare.

> „*Se focusează pe un lucru și repetă verbal acel lucru.”*

Figura 6.12 Răspunsul participantului P_{10} la întrebarea Q_3 din testul DISCOVER.

> „*Comunicarea se realizează dificil cu cuvinte fragmentate și greu inteligibile de o terță persoană. De multe ori nu creează contact vizual când i se solicită un răspuns, dar reacționează bine la prompt kinetic.”*

Figura 6.13 Răspunsul participantului P_{22} la întrebarea Q_2 din testul DISCOVER.

Figura 6.13 prezintă răspunsul participantului P_{22} la întrebarea Q_2. Modelul de AI a identificat cu succes simptomele TSA „deficiențe de comunicare” și „contact vizual redus”. Rezultatul pozitiv se datorează în parte preciziei ridicate a clasei „deficiențe de comunicare” (p=84.62%) și răspunsului elaborat al părintelui P_{22}.

7 CONCLUZII

Text mining este o paradigmă contemporană ce se concentrează pe analiza computerizată a cantități mari de date text. Într-o epocă informațională, în care datele joacă un rol central în modelarea peisajelor socio-economice, această teză investighează potențialul neexplorat al text mining și prezintă o abordare de pionierat prin propunerea utilizării text mining ca tehnologie de sprijin a deciziilor de diagnosticare a autismului.

Pentru realizarea acestui obiectiv s-a abordat o metodologie centrată pe om pentru a înțelege nevoile pacienților și a specialiștilor medicali și modalitatea în care tehnologia le poate aborda. În acest sens, au fost parcurse următoarele etape de cercetare: (1) analiza stadiului actual în domeniul text mining, care a implicat investigarea și sintetizarea metodelor și tehnicilor de procesare a textului descrise în literatura de specialitate și practicate de comunitatea științifică; (2) prezentarea conceptului clinical text mining și a beneficiilor pe care le poate aduce în domeniul screening-ului TSA comparativ cu metodele contemporane, identificate prin revizuirea articolelor relevante din bazele de date PubMed și ACM Digital Library; (3) analiza empirică a percepției părinților și specialiștilor medicali față de diagnosticarea autismului asistată de tehnologie, care a constat într-un experiment controlat efectuat pe un eșantion de N=56 de participanți distribuiți în două grupuri; (4) proiectarea și implementarea a două prototipuri de tehnologie, care să identifice simptome TSA în date text nestructurate și examinarea performanței modelelor AI antrenate; (5) analiza empirică a percepției părinților cu copil cu diagnostic TSA și a părinților cu copil fără diagnostic TSA față de diagnosticarea autismului asistată de prototipul Autism AI Advisor For Diagnosis, care a constat într-un experiment controlat efectuat pe un eșantion de N=22 de participanți distribuiți în două grupuri.

Astfel, această teză a raportat rezultate privind percepțiile utilizatorilor față de diagnosticarea TSA asistată de tehnologie. Constatările au subliniat relații relevante între variabilele socio-demografice, variabilele care descriu comportamentul de utilizare a tehnologiilor moderne și fezabilitatea percepută de participanți față de diagnosticarea autismului asistată de tehnologie. În general, rezultatele au arătat o atitudine pozitivă față de diagnosticarea TSA mediată de tehnologie. Importanța atribuită de eșantion pentru tehnologiile de diagnosticare a fost „foarte mare” și „mare”, profesioniștii din domeniul sănătății văzând un potențial mai mare decât părinții.

În teză au fost prezentate procesele de proiectare și implementare a două prototipuri de sisteme de suport pentru deciziile de diagnosticare TSA, bazate pe tehnologia de text minig. Dezvoltarea prototipurilor a început cu crearea unui corpus de date, printr-un experiment controlat efectuat pe un eșantion de N=44 de participanți. Sarcina participanților a fost să răspundă la un test de screening, conceput special pentru a permite răspunsuri în format liber. Corpusul a fost supus procesului KDT în vederea antrenării modelelor de AI. Obiectivul modelelor a fost identificarea de tipare și indicatori lingvistici care pot contribui la detectarea timpurie a simptomelor TSA. Performanța prototipului dezvoltat în RapidMiner a fost analizată comparativ prin utilizarea algoritmilor Naïve Bayes, K-Nearest Neighbors, Deep Learning și Radom Forest. Modelul antrenat folosind k-NN a produs cea mai mare acuratețe de 78.69%. Rezultatul prototipului Autism AI Advisor for Diasgnosis, dezvoltat folosind Microsoft Language Studio, a demonstrat o precizie de 91.57%. Rezultatele analizei de performanță sunt corelate cu implementările tehnice. Primul prototip abordează clasificarea cu etichetă unică, iar al doilea prototip implementează clasificarea cu etichete multiple. În primul prototip a fost introdusă o etapă de traducere a datelor din limba română în limba engleză, în timp ce al doilea utilizează datele în limba română, fără a altera într-un fel sensul mesajului transmis.

Rezultatele performante a prototipului Autism AI Advisor For Diagnosis au încurajat explorarea utilizabilității aplicației print-un experiment controlat cu N=22 de participanți. Metricile folosite pentru evaluare au fost scalele CSE și SUS, instrumentul APSI și chestionarul

de percepție a fezabilității aplicației. Rezultatele au indicat un nivel bun de utilizabilitate percepută SUS=81.36% și o percepție pozitivă, participanții raportând un nivel înalt de mulțumire și intenția de a recomanda instrumentul de screening și altor persoane.

Cu toate acestea, rezultatele necesită o reflecție profundă asupra limitărilor. Participanții la studii au fost recrutați preponderent din România, ceea ce poate duce la rezultate influențate de aspectele sociale și culturale specifice acestei țări. Cercetările ulterioare ar putea implica persoane cu medii culturale și sociale diverse pentru a crea un corpus de date mai cuprinzător, ceea ce are potențialul de a îmbogăți rezultatele. De asemenea, cercetările viitoare ar putea explora și abordări inovatoare precum convertirea vorbirii în text (*en.*: speech to text). Implementarea unui sistem de suport de decizii pentru diagnosticarea TSA, care în etapa de screening elimină necesitatea de a scrie, prin procesarea facilă a sursei audio pentru transcrierea rapidă și precisă a vorbirii, ar putea contribui la îmbunătățirea performanței modelelor AI. Textul astfel procesat ar conține mai puține erori gramaticale. În plus, o astfel de opțiune ar face accesibilă tehnologia și pentru persoane cu deficiențe motorii fine.

PROIECTE DE CERCETARE

Această teză a fost susținută și de următoarele proiecte de cercetare:

- „Excelență academică și valori antreprenoriale - sistem de burse pentru asigurarea oportunităților de formare și dezvoltare a competențelor antreprenoriale ale doctoranzilor și post doctoranzilor - ANTREPRENORDOC”,
  Contract de finanțare nr. 36355/23.05.2019 POCU/380/6/13
  Contract subsidiar nr. SMIS 123847.

- „Centru pentru transferul de cunoștințe către întreprinderi din domeniul ICT—CENTRIC, Autism ASSISTant - Asistent virtual pentru dezvoltarea abilităților cognitive ale copiilor cu patologie de spectru autist”,
  Contract de finanțare nr. 5/AXA 1/1.2.3/G/13.06.2018
  Contract subsidiar nr. 22080/05.10.2022/Autism ASSISTant/ASSIST.

- „Centru pentru transferul de cunoștințe către întreprinderi din domeniul ICT—CENTRIC, PARS - Platforma Autonomă de Recunoaștere și Suport”
  Contract de finanțare nr. 5/AXA 1/1.2.3/G/13.06.2018
  Contract subsidiar nr. 22081/05.10.2022/PARS/ASSIST.

REFERINȚE BIBLIOGRAFICE

[1] A. F. A. H. Alnuaimi and T. H. K. Albaldawi, "An overview of machine learning classification techniques," *BIO Web Conf.*, vol. 97, p. 00133, 2024, doi: 10.1051/bioconf/20249700133.

[2] M. G. H. Omran, A. P. Engelbrecht, and A. Salman, "An overview of clustering methods," *IDA*, vol. 11, no. 6, pp. 583–605, Nov. 2007, doi: 10.3233/IDA-2007-11602.

[3] T. Barnett, "The Zettabyte Era Officially Begins (How Much is That?," Cisco Blogs.

[4] P. Michel, *The Use of Technology in the Study, Diagnosis and Treatment of Autism*. 2004.

[5] M. Roser and H. Ritchie, "How has world population growth changed over time?," *Our World in Data*, 2023.

[6] P. Taylor, "The amount of data created, consumed, and stored 2010-2020, with forecasts to 2025." [Online]. Available: https://www.statista.com/statistics/871513/%20worldwide-data-created

[7] T. King, "80 Percent of Your Data Will Be Unstructured in Five Years," Data Management Solutions Review.

[8] *Working with text: tools, techniques and approaches for text mining*, 1st edition. Waltham, MA: Elsevier, 2016.

[9] Data Bridge Market Research, "Global Text Analytics Market – Industry Trends and Forecast to 2029." [Online]. Available: https://www.databridgemarketresearch.com/reports/global-text-analytics-market

[10] K. Warwick and H. Shah, "Can machines think? A report on Turing test experiments at the Royal Society," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 28, no. 6, pp. 989–1007, Nov. 2016, doi: 10.1080/0952813X.2015.1055826.

[11] "EU Digital Strategy." [Online]. Available: https://eufordigital.eu/discover-eu/eu-digital-strategy/

[12] "Deceniul digital al Europei: obiective digitale pentru 2030." [Online]. Available: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_ro

[13] *2030 Digital Compass: the European way for the Digital Decade*. Brussels, 2021.

[14] "Patient safety." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/patient-safety

[15] Petruta Teampau, "Introducere în metodologia cercetării științelor sociale."

[16] *HORIZON 2020 – Work Programme 2018-2020, General Annexes*. 2019.

[17] F. M. D. A. Dias, D. F. Rodrigues, and C. H. M. D. Souza, "Mobile Applications for autisticchildren: An analysis of the Google Play Store platform," *IJAERS*, vol. 7, no. 9, pp. 476–486, 2020, doi: 10.22161/ijaers.79.55.

[18] G. Miner, *Practical text mining and statistical analysis for non-structured text data applications*, 1st ed. Waltham, MA: Academic Press, 2012.

[19] R. Feldman, Ronen, Sanger, and James, *The text mining handbook: Advanced approaches in analyzing unstructured data*. 2007.

[20] M. Hearst, "What Is Text Mining?" [Online]. Available: https://people.ischool.berkeley.edu/~hearst/text-mining.html

[21] H. Yan, M. Ma, Y. Wu, H. Fan, and C. Dong, "Overview and analysis of the text mining applications in the construction industry," *Heliyon*, vol. 8, no. 12, p. e12088, Dec. 2022, doi: 10.1016/j.heliyon.2022.e12088.

[22] R. Feldman and I. Dagan, "Knowledge Discovery in Textual Databases (KDT)," Jun. 1995.

[23] M. Agosti and A. F. Smeaton, *Information Retrieval and Hypertext*. Boston, MA: Springer US, 1996.

[24] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organizational Research Methods*, vol. 25, no. 1, pp. 114–146, Jan. 2022, doi: 10.1177/1094428120971683.

[25] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *IJCA*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.

[26] Q. Chen and M. Sokolova, "Specialists, Scientists, and Sentiments: Word2Vec and Doc2Vec in Analysis of Scientific and Medical Texts," *SN COMPUT. SCI.*, vol. 2, no. 5, p. 414, Sep. 2021, doi: 10.1007/s42979-021-00807-1.

[27] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[28] W. B. Zulfikar, M. Irfan, C. N. Alam, and M. Indra, "The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter," in *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, Denpasar, Bali, Indonesia: IEEE, Aug. 2017, pp. 1–5. doi: 10.1109/CITSM.2017.8089231.

[29] L. Wei, B. Wei, and B. Wang, "Text Classification Using Support Vector Machine with Mixture of Kernel," *JSEA*, vol. 05, no. 12, pp. 55–58, 2012, doi: 10.4236/jsea.2012.512B012.

[30] J. Sun, W. Du, and N. Shi, "A Survey of kNN Algorithm," *Inf Eng Appl Comput*, vol. 1, no. 1, May 2018, doi: 10.18063/ieac.v1i1.770.

[31] A. Shrestha and A. Mahmood, "Review of Deep Learning Algorithms and Architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019, doi: 10.1109/ACCESS.2019.2912200.

[32] A. Aliero, S. Bashir, H. Aliyu, A. Tafida, B. Kangiwa, and N. Dankolo, "Systematic Review on Text Normalization Techniques and its Approach to Non-Standard Words," *International Journal of Computer Applications*, vol. 185, pp. 975–8887, Sep. 2023.

[33] V. Mohan, “Text Mining: Open Source Tokenization Tools: An Analysis,” vol. 3, pp. 37–47, Jan. 2016.

[34] M. Lamba and M. Madhusudhan, *Text mining for information professionals: an uncharted territory*. Cham: Springer, 2022.

[35] Department of Computer Science and Engineering Sant Longowal Institute of Engineering and Technology Longowal, Sangrur (Punjab) – 148106, India and J. Kaur, “STOPWORDS REMOVAL AND ITS ALGORITHMS BASED ON DIFFERENT METHODS,” *ijarcs*, vol. 9, no. 5, pp. 81–88, Oct. 2018, doi: 10.26483/ijarcs.v9i5.6301.

[36] N. Indurkhya and F. J. Damerau, *Handbook of natural language processing*. Boca Raton, FL: Chapman & Hall/CRC, 2010.

[37] J. B. Lovins, “Development of a Stemming Algorithm,” *Mechanical Translation and Computational Linguistics*, vol. 11, 1968.

[38] M. F. Porter, “The Porter Stemming Algorithm,” 2006, [Online]. Available: https://tartarus.org/martin/PorterStemmer/index.html

[39] N. A. Razmi, M. Z. Zamri, S. S. S. Ghazalli, and N. Seman, “Visualizing stemming techniques on online news articles text analytics,” *Bulletin EEI*, vol. 10, no. 1, pp. 365–373, Feb. 2021, doi: 10.11591/eei.v10i1.2504.

[40] M. F. Porter, “Snowball: A language for stemming algorithms,” 2001, [Online]. Available: http://snowball.tartarus.org/texts/introduction

[41] M. Mhiri, C. Desrosiers, and M. Cheriet, “Word spotting and recognition via a joint deep embedding of image and text,” *Pattern Recognition*, vol. 88, pp. 312–320, Apr. 2019, doi: 10.1016/j.patcog.2018.11.017.

[42] Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, and Brown, “Text Classification Algorithms: A Survey,” *Information*, vol. 10, no. 4, p. 150, Apr. 2019, doi: 10.3390/info10040150.

[43] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, “Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach,” in *2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)*, Yogyakarta, Indonesia: IEEE, Oct. 2014, pp. 1–4. doi: 10.1109/ICITEED.2014.7007894.

[44] M. Mouriño-García, R. Perez-Rodriguez, and L. Anido-Rifón, “Bag-of-Concepts Document Representation for Textual News Classification,” *International Journal of Computational Linguistics and Applications*, vol. 6, pp. 173–188, Jun. 2015.

[45] J. A. Nichols, H. W. Herbert Chan, and M. A. B. Baker, “Machine learning: applications of artificial intelligence to imaging and diagnosis,” *Biophys Rev*, vol. 11, no. 1, pp. 111–118, Feb. 2019, doi: 10.1007/s12551-018-0449-9.

[46] J. Žižka, F. Dařena, and A. Svoboda, *Text Mining with Machine Learning: Principles and Techniques*, 1st ed. First. | Boca Raton : CRC Press, 2019.: CRC Press, 2019. doi: 10.1201/9780429469275.

[47] Q. Liu and Y. Wu, "Supervised Learning," in *Encyclopedia of the Sciences of Learning*, N. M. Seel, Ed., Boston, MA: Springer US, 2012, pp. 3243–3245. doi: 10.1007/978-1-4419-1428-6_451.

[48] M. Allahyari *et al.*, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," Jul. 2017.

[49] M. W. Berry and M. Castellanos, Eds., *Survey of Text Mining II*. London: Springer London, 2008. doi: 10.1007/978-1-84800-046-9.

[50] N. Jokar, A. R. Honarvar, S. Aghamirzadeh, and K. Esfandiari, "Web mining and Web usage mining techniques," *Bull. Soc. Roy. Sc. de Liège*, pp. 321–328, 2016, doi: 10.25518/0037-9565.5371.

[51] R. Kibble, *Introduction to natural language processing*. University of London, 2013.

[52] C. Luque, J. M. Luna, M. Luque, and S. Ventura, "An advanced review on text mining in medicine," *WIREs Data Min & Knowl*, vol. 9, no. 3, p. e1302, May 2019, doi: 10.1002/widm.1302.

[53] H. Dalianis, *Clinical text mining: secondary use of electronic patient records*. Cham, Switherland: Springer, 2018.

[54] A. I. Fedotchev, V. V. Dvoryaninova, S. D. Velikova, and A. A. Zemlyanaya, "Modern Technologies in Studying the Mechanisms, Diagnostics, and Treatment of Autism Spectrum Disorders (Review)," *Sovrem Tehnol Med*, vol. 11, no. 1, p. 31, Mar. 2019, doi: 10.17691/stm2019.11.1.03.

[55] J. N. Constantino, "Deconstructing autism: from unitary syndrome to contributory developmental endophenotypes," *International Review of Psychiatry*, vol. 30, no. 1, pp. 18–24, Jan. 2018, doi: 10.1080/09540261.2018.1433133.

[56] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition. American Psychiatric Association, 2013. doi: 10.1176/appi.books.9780890425596.

[57] Lecturer, PhD, Department of Educational Sciences, The University Clinic for Therapies and PsychoPedagogical Counselling, West University of Timișoara, loredana.al@e-uvt.ro and L. Al Ghazi, "History of autism. The beginnings. Collusions or serendipity," *JES*, vol. 38, no. 2, pp. 5–17, Dec. 2018, doi: 10.35923/JES.2018.2.01.

[58] J. P. T. Higgins *et al.*, Eds., *Cochrane Handbook for Systematic Reviews of Interventions*, 1st ed. Wiley, 2019. doi: 10.1002/9781119536604.

[59] R. Nur Syahindah Husna, A. R. Syafeeza, N. Abdul Hamid, Y. C. Wong, and R. Atikah Raihan, "FUNCTIONAL MAGNETIC RESONANCE IMAGING FOR AUTISM SPECTRUM DISORDER DETECTION USING DEEP LEARNING," *Jurnal Teknologi*, vol. 83, no. 3, pp. 45–52, Apr. 2021, doi: 10.11113/jurnalteknologi.v83.16389.

[60] R. S. Chivate *et al.*, "PET/CT in Autism, A Diagnostic tool?," 2016, doi: 10.13140/RG.2.1.2080.5525.

[61] M. Milovanovic and R. Grujicic, "Electroencephalography in Assessment of Autism Spectrum Disorders: A Review," *Front. Psychiatry*, vol. 12, p. 686021, Sep. 2021, doi: 10.3389/fpsyt.2021.686021.

[62] A. Kouroupa, K. R. Laws, K. Irvine, S. E. Mengoni, A. Baird, and S. Sharma, "The use of social robots with children and young people on the autism spectrum: A systematic review and meta-analysis," *PLoS ONE*, vol. 17, no. 6, p. e0269800, Jun. 2022, doi: 10.1371/journal.pone.0269800.

[63] M. Briguglio *et al.*, "A Machine Learning Approach to the Diagnosis of Autism Spectrum Disorder and Multi-Systemic Developmental Disorder Based on Retrospective Data and ADOS-2 Score," *Brain Sciences*, vol. 13, no. 6, p. 883, May 2023, doi: 10.3390/brainsci13060883.

[64] A. V. Dahiya, E. DeLucia, C. G. McDonnell, and A. Scarpa, "A systematic review of technological approaches for autism spectrum disorder assessment in children: Implications for the COVID-19 pandemic," *Research in Developmental Disabilities*, vol. 109, p. 103852, Feb. 2021, doi: 10.1016/j.ridd.2021.103852.

[65] P. Defresne and L. Mottron, "Clinical Situations in Which the Diagnosis of Autism is Debatable: An Analysis and Recommendations," *Can J Psychiatry*, vol. 67, no. 5, pp. 331–335, May 2022, doi: 10.1177/07067437211041469.

[66] V. Guinchat *et al.*, "Very early signs of autism reported by parents include many concerns not specific to autism criteria," *Research in Autism Spectrum Disorders*, vol. 6, no. 2, pp. 589–601, Apr. 2012, doi: 10.1016/j.rasd.2011.10.005.

[67] S. Siklos and K. A. Kerns, "Assessing the diagnostic experiences of a small sample of parents of children with autism spectrum disorders," *Research in Developmental Disabilities*, vol. 28, no. 1, pp. 9–22, Jan. 2007, doi: 10.1016/j.ridd.2005.09.003.

[68] L. Klintwall, S. Eldevik, and S. Eikeseth, "Narrowing the gap: Effects of intervention on developmental trajectories in autism," *Autism*, vol. 19, no. 1, pp. 53–63, Jan. 2015, doi: 10.1177/1362361313510067.

[69] O. Akinnusotu, A. Bhatti, C. A. Doubeni, and M. Williams, "Supporting Mental Health and Psychological Resilience Among the Health Care Workforce: Gaps in the Evidence and Urgency for Action," *Ann Fam Med*, vol. 21, no. Suppl 2, pp. S100–S102, Feb. 2023, doi: 10.1370/afm.2933.

[70] L. Herlihy, K. Knoch, B. Vibert, and D. Fein, "Parents' first concerns about toddlers with autism spectrum disorder: Effect of sibling status," *Autism*, vol. 19, no. 1, pp. 20–28, Jan. 2015, doi: 10.1177/1362361313509731.

[71] M. Regalado and N. Halfon, "Primary Care Services Promoting Optimal Child Development From Birth to Age 3 Years: Review of the Literature," *Arch Pediatr Adolesc Med*, vol. 155, no. 12, p. 1311, Dec. 2001, doi: 10.1001/archpedi.155.12.1311.

[72] F. P. Glascoe, "Parents' Evaluation of Developmental Status: How Well Do Parents' Concerns Identify Children With Behavioral and Emotional Problems?," *Clin Pediatr (Phila)*, vol. 42, no. 2, pp. 133–138, Mar. 2003, doi: 10.1177/000992280304200206.

[73] C. Skellern, Y. Rogers, and M. O'Callaghan, "A parent-completed developmental questionnaire: Follow up of ex-premature infants," *J Paediatrics Child Health*, vol. 37, no. 2, pp. 125–129, Apr. 2001, doi: 10.1046/j.1440-1754.2001.00604.x.

[74] U. Raja, T. Mitchell, T. Day, and J. M. Hardin, “Text mining in healthcare. Applications and opportunities.,” *J Healthc Inf Manag*, vol. 22, no. 3, pp. 52–56, Summer 2008.

[75] P. Nitiéma, “Artificial Intelligence in Medicine: Text Mining of Health Care Workers’ Opinions,” *J Med Internet Res*, vol. 25, p. e41138, Jan. 2023, doi: 10.2196/41138.

[76] I. Hendrickx, T. Voets, P. Van Dyk, and R. B. Kool, “Using Text Mining Techniques to Identify Health Care Providers With Patient Safety Problems: Exploratory Study,” *J Med Internet Res*, vol. 23, no. 7, p. e19064, Jul. 2021, doi: 10.2196/19064.

[77] V. Pendyala, Fang, J. Holliday, and A. Zalzala, *A text mining approach to automated healthcare for the masses*. 2014. doi: 10.1109/GHTC.2014.6970257.

[78] D. L. Robins, D. Fein, M. L. Barton, and J. A. Green, “Modified Checklist for Autism in Toddlers.” May 07, 2012. doi: 10.1037/t03999-000.

[79] M. Shucksmith, S. Cameron, T. Merridew, and F. Pichler, “Urban–Rural Differences in Quality of Life across the European Union,” *Regional Studies*, vol. 43, no. 10, pp. 1275–1289, Dec. 2009, doi: 10.1080/00343400802378750.

[80] E. D. Hacker, “Technology and quality of life outcomes,” *Semin Oncol Nurs*, vol. 26, no. 1, pp. 47–58, Feb. 2010, doi: 10.1016/j.soncn.2009.11.007.

[81] R. Sulek *et al.*, “Support Preferences and Clinical Decision Support Systems (CDSS) in the Clinical Care of Autistic Children: Stakeholder Perspectives,” *Adv Neurodev Disord*, Jul. 2024, doi: 10.1007/s41252-024-00410-4.

[82] A. T. M. Wasylewicz and A. M. J. W. Scheepers-Hoeks, “Clinical Decision Support Systems,” in *Fundamentals of Clinical Data Science*, P. Kubben, M. Dumontier, and A. Dekker, Eds., Cham: Springer International Publishing, 2019, pp. 153–169. doi: 10.1007/978-3-319-99713-1_11.

[83] *What is an AI model?* [Online]. Available: https://www.ibm.com/topics/ai-model

[84] F. J. W. M. Dankers, A. Traverso, L. Wee, and S. M. J. Van Kuijk, “Prediction Modeling Methodology,” in *Fundamentals of Clinical Data Science*, P. Kubben, M. Dumontier, and A. Dekker, Eds., Cham: Springer International Publishing, 2019, pp. 101–120. doi: 10.1007/978-3-319-99713-1_8.

[85] N. Kassim and H. Hashim, “Common European Framework of Reference (CEFR): A Review on its Implementation in ESL/EFL Classrooms,” *IJARBSS*, vol. 13, no. 12, p. Pages 2991-3016, Dec. 2023, doi: 10.6007/IJARBSS/v13-i12/20149.

[86] T. C. McFayden, O. Putnam, R. Grzadzinski, and C. Harrop, “Sex Differences in the Developmental Trajectories of Autism Spectrum Disorder,” *Curr Dev Disord Rep*, vol. 10, no. 1, pp. 80–91, Jan. 2023, doi: 10.1007/s40474-023-00270-y.

[87] American Psychiatric Association and American Psychiatric Association, Eds., *Diagnostic and statistical manual of mental disorders: DSM-5*, 5th ed. Washington, D.C: American Psychiatric Association, 2013.

[88] S. J. Webb and E. J. H. Jones, “Early Identification of Autism: Early Characteristics, Onset of Symptoms, and Diagnostic Stability,” *Infants & Young Children*, vol. 22, no. 2, pp. 100–118, Apr. 2009, doi: 10.1097/IYC.0b013e3181a02f7f.

[89] G. Ertek, D. Tapucu, and I. Arın, *Text Mining with RapidMiner*, vol. Markus Hofmann, Ralf Klinkenberg (Eds.) RapidMiner: Data Mining Use Cases and Business Analytics Applications. Chapman&Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC. 2013.

[90] T. A. Mat, A. Lajis, and H. Nasir, "Text Data Preparation in RapidMiner for Short Free Text Answer in Assisted Assessment," in *2018 IEEE 5th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)*, Songkla, Thailand: IEEE, Nov. 2018, pp. 1–4. doi: 10.1109/ICSIMA.2018.8688806.

[91] J. Singh and V. Gupta, "Text Stemming: Approaches, Applications, and Challenges," *ACM Comput. Surv.*, vol. 49, no. 3, pp. 1–46, Sep. 2017, doi: 10.1145/2975608.

[92] V. Mallawaarachchi, "Porter stemming algorithm - basic intro."

[93] F.-J. Yang, "An Implementation of Naive Bayes Classifier," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA: IEEE, Dec. 2018, pp. 301–306. doi: 10.1109/CSCI46756.2018.00065.

[94] "K-Nearest Neighbor(KNN) Algorithm." [Online]. Available: https://www.geeksforgeeks.org/k-nearest-neighbours/

[95] "RapidMiner Documentation." [Online]. Available: https://docs.rapidminer.com

[96] S. Wang, C. Aggarwal, and H. Liu, "Random-Forest-Inspired Neural Networks," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 6, pp. 1–25, Nov. 2018, doi: 10.1145/3232230.

[97] A. Balasch, M. Beinhofer, and G. Zauner, "The Relative Confusion Matrix, a Tool to Assess Classifiablility in Large Scale Picking Applications," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, Paris, France: IEEE, May 2020, pp. 8390–8396. doi: 10.1109/ICRA40945.2020.9197540.

[98] E. Gordon-Lipkin, J. Foster, and G. Peacock, "Whittling Down the Wait Time," *Pediatric Clinics of North America*, vol. 63, no. 5, pp. 851–859, Oct. 2016, doi: 10.1016/j.pcl.2016.06.007.

[99] M. Penner, E. Anagnostou, and W. J. Ungar, "Practice patterns and determinants of wait time for autism spectrum disorder diagnosis in Canada," *Molecular Autism*, vol. 9, no. 1, p. 16, Dec. 2018, doi: 10.1186/s13229-018-0201-0.

[100] M. Stoica, "Telemedicina si e-sanatate," *Informatica Economica*, vol. 1, no. 21.

[101] P. Kubben, "Mobile Apps," in *Fundamentals of Clinical Data Science*, P. Kubben, M. Dumontier, and A. Dekker, Eds., Cham: Springer International Publishing, 2019, pp. 171–179. doi: 10.1007/978-3-319-99713-1_12.

[102] N. Babich, "Human-Centered Design: An Introduction, Practices, and Principles." [Online]. Available: https://www.shopify.com/partners/blog/human-centered-design

[103] M. Melles, A. Albayrak, and R. Goossens, "Innovating health care: key characteristics of human-centered design," *Int J Qual Health Care*, vol. 33, no. Supplement_1, pp. 37–44, Jan. 2021, doi: 10.1093/intqhc/mzaa127.

[104] G. E. Constain Moreno, C. A. Collazos, S. B. Blasco, and F. Moreira, "Software Design for Users with Autism Using Human-Centered Design and Design Thinking

Techniques," *Sustainability*, vol. 15, no. 24, p. 16587, Dec. 2023, doi: 10.3390/su152416587.

[105] D. Nessler, "How to apply a design thinking, HCD, UX or any creative process from scratch — Revised & New Version." [Online]. Available: https://uxdesign.cc/how-to-solve-problems-applying-a-uxdesign-designthinking-hcd-or-any-design-process-from-scratch-v2-aa16e2dd550b

[106] A. J. Jacobs, "UX: Creating Proto-Personas." [Online]. Available: https://uxdesign.cc/ux-creating-proto-personas-76a1738401a2

[107] M. F. Santoso, "Implementation Of UI/UX Concepts And Techniques In Web Layout Design With Figma," *JTEKSIS*, vol. 6, no. 2, pp. 279–285, Apr. 2024, doi: 10.47233/jteksis.v6i2.1223.

[108] C. Allison *et al.*, "The Q-CHAT (Quantitative CHecklist for Autism in Toddlers): A Normally Distributed Quantitative Measure of Autistic Traits at 18–24 Months of Age: Preliminary Report," *J Autism Dev Disord*, vol. 38, no. 8, pp. 1414–1425, Sep. 2008, doi: 10.1007/s10803-007-0509-7.

[109] "Firebase Realtime Database." [Online]. Available: https://firebase.google.com/docs/database

[110] "What is Azure AI Language?" [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-services/language-service/overview

[111] C.-H. Tsai *et al.*, "The symptoms of autism including social communication deficits and repetitive and restricted behaviors are associated with different emotional and behavioral problems," *Sci Rep*, vol. 10, no. 1, p. 20509, Nov. 2020, doi: 10.1038/s41598-020-76292-y.

[112] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation," in *Advances in Information Retrieval*, vol. 3408, D. E. Losada and J. M. Fernández-Luna, Eds., in Lecture Notes in Computer Science, vol. 3408. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 345–359. doi: 10.1007/978-3-540-31865-1_25.

[113] R. Loomes, L. Hull, and W. P. L. Mandy, "What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 56, no. 6, pp. 466–474, Jun. 2017, doi: 10.1016/j.jaac.2017.03.013.

[114] M. C. Howard, "Creation of a Computer Self-Efficacy Measure: Analysis of Internal Consistency, Psychometric Properties, and Validity," *Cyberpsychology, Behavior, and Social Networking*, vol. 17, no. 10, pp. 677–681, Oct. 2014, doi: 10.1089/cyber.2014.0255.

[115] L. M. T. Silva and M. Schalock, "Autism Parenting Stress Index: Initial Psychometric Evidence," *J Autism Dev Disord*, vol. 42, no. 4, pp. 566–574, Apr. 2012, doi: 10.1007/s10803-011-1274-1.

[116] V. Venkatesh, "Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model," *Information Systems Research*, vol. 11, no. 4, pp. 342–365, Dec. 2000, doi: 10.1287/isre.11.4.342.11872.

[117] Ping Zhang, Na Li, and Heshan Sun, "Affective Quality and Cognitive Absorption: Extending Technology Acceptance Research," in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, Kauia, HI, USA: IEEE, 2006, pp. 207a–207a. doi: 10.1109/HICSS.2006.39.

[118] J. Brooke, "SUS -- a quick and dirty usability scale," 1996, pp. 189–194.

[119] A. Bangor, P. Kortum, and J. Miller, "Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale," *J. Usability Stud.*, vol. 4, pp. 114–123, Apr. 2009.

[120] F. Xie, E. Pascual, and T. Oakley, "Functional echolalia in autism speech: Verbal formulae and repeated prior utterances as communicative and cognitive strategies," *Front. Psychol.*, vol. 14, p. 1010615, Feb. 2023, doi: 10.3389/fpsyg.2023.1010615.