”ALGORITMI SI UN SET DE DATE CARE FOLOSESC DINAMICA TASTĂRII TASTELOR IN CAZUL TEXTULUI SCRIS LIBER PENTRU AUTENTIFICAREA CONTINUĂ IN PLATFORME EDUCATIONALE CARE CUPRIND CUSRURI ONLINE DESCHISE MASIVE (MOOC)” /

”FREE-TEXT KEYSTROKE DYNAMICS DATA SET AND ALGORITHMS FOR CONTINUOUS AUTHENTICATION IN EDUCATIONAL PLATFORMS WITH MASSIVE OPEN ONLINE COURSES (MOOC)”

Teză de doctorat – Rezumat

pentru obținerea titlului științific de doctor la

Universitatea Politehnica Timișoara

în domeniul de doctorat Calculatoare și Tehnologia Informației

autor ing. Augustin-Cătălin IAPĂ

conducător științific Prof.univ.emerit dr.ing. Vladimir-Ioan CREȚU

februarie 2021

Cuprins:

ALGORITMI SI UN SET DE DATE CARE FOLOSESC DINAMICA TASTĂRII TASTELOR IN CAZUL TEXTULUI SCRIS LIBER PENTRU AUTENTIFICAREA CONTINUĂ IN PLATFORME EDUCATIONALE CARE CUPRIND CUSRURI ONLINE DESCHISE MASIVE (MOOC)

Prezenta lucrare se focusează pe autentificarea continuă a utilizatorului unui calculator pe baza modului de a tasta la tastatură. În cadrul cercetării s-a dezvoltat un algoritm de autentificare pe baza modului de tastare, s-a colectat un set de date referitoare la modul de tastare de la 80 de voluntari, s-au propus două metrici modificate pentru a se obține performanțe mai bune ale algoritmului de autentificare și s-a propus o structură de date pentru

a stoca informațiile necesare ale utilizatorilor.

Această metodă de autentificare își justifică atenția mai ales în cadrul platformelor educaționale online, platforme care au cunoscut o creștere foarte mare în anul 2020, datorită mutării majorității cursurilor în mediul online, restricție generată de criza COVID-19.

1. INTRODUCERE

Prezenta lucrare a pornit de la necesitatea dezvoltării unor modalități suplimentare de a identifica identitatea unui utilizator care folosește un cont privat pe un calculator. Aceasta nevoie e mai pronunțata in cazul cursurilor care se desfășoară on-line sau in cazul examenelor pe care studenții sau elevii le susțin in sisteme educaționale on-line. Începând cu anul 2008 a luat naștere fenomenul MOOC (Masive Open Online Courses), adică cursuri la care participa un număr mare de studenți din orice colt al lumii in sistem online. Acest fenomen a atins un prim maxim in anul 2012, iar in anul 2020 a cunoscut o creștere exponențiala a numărului de studenți înrolați.

Anul 2020 a dus, de asemenea, la creșterea radicala a utilizării sistemelor educaționale on-line in contextul crizei sanitare provocate de virusul SARS-CoV-2. Universități, scoli primare sau licee au fost obligate sa se adapteze si sa mute întreg sistemul educațional clasic, in clasa, fata in fata pe platforme la distanta. In aceste contexte a devenit mult mai importanta găsirea de metode pentru asigurarea ca la un examen, la care atât profesorul cat si studenții sunt in locații diferite, prin intermediul unor mijloace ușor accesibile sa ne asiguram ca studentul este cel care rezolva subiectele si primește nota pe cunoștințele sale.

Exista numeroase metode de a identifica si autentifica un utilizator într-un cont electronic. Cea mai răspândita metoda este reținerea unui utilizator si a parolei si pe baza celor doua utilizatorul are acces la cont. Folosirea unor carduri fizice, cum sunt cele folosite de bănci, sau a amprentei, scanării retinei sau recunoaștere facila presupune existenta unor dispozitive suplimentare pentru preluarea de date de la utilizatori. Pentru autentificarea in timpul unui examen nu e suficient sa existe un cont si o parola, in cazul in care studentul vrea sa lase pe altcineva in locul lui pentru a rezolva subiectele. Cele mai multe examene se dau acum cu camera de filmat si microfonul pornite pe tot parcursul examenului.

O metoda eficienta in rezolvarea problemei descrise mai sus o reprezintă autentificarea continua cu ajutorul keystroke dynamics. Keystroke dynamics este metoda prin care un utilizator poate fi identificat sau autentificat pe baza modului sau particular de a tasta un text la tastatura. Aceasta metoda nu necesita hardware suplimentar, orice calculator sau laptop este dotat cu o tastatura. De asemenea un alt avantaj este reprezentat de faptul ca verificarea identității se poate face continuu, in orice moment in care utilizatorul scrie la tastatura. Autentificarea cu parola nu se poate face continuu, făcând-se, de regula o singura data la accesarea contului, iar pe parcurs putând sa se schimbe utilizatorul fără ca sistemul sa își dea seama de schimbare.

Un alt avantaj al utilizării identificării sau autentificării cu ajutorul keystroke dynamics este acela ca utilizatorul nu are de făcut pași suplimentari. Pur si simplu trebuie sa scrie si sistemul monitorizează modul de a tasta. In acest caz, după o autentificare într-un sistem, daca se schimba utilizatorul, sistemul își va da seama ca altcineva este la calculator si poate semnala aceasta schimbare.

La cursuri MOOC pot participa mii de studenți in același timp. In cazul unui examen cu mii de studenți devine imposibila supravegherea prin intermediul camerei video si a microfonului, aceasta metoda fiind eficienta când numărul de studenți este mic. In cazul keystroke dynamics pot fi autentificați in mod continuu orice număr de studenți, oricât de mare ar fi acesta, nu exista aceasta limitare.

Dezavantajul unui sistem cu autentificare sau identificare a utilizatorilor prin

intermediul metodei keystroke dynamics îl reprezintă exactitatea algoritmului cu care se poate face identificarea utilizatorului. In prezent, sistemele care folosesc aceasta metoda nu ating rate de eroare de 0%. Au performate care identifica utilizatorul cu un procent de eroare de sub 10%, sau in unele cazuri si cu exactitate mai mare, in schimb îmbunătățirea algoritmilor bazați pe keystroke dynamics reprezintă in continuare o provocare in acest domeniu. O alta provocare pentru cercetarea științifica in acest domeniu o reprezintă faptul ca pentru a testa eficienta algoritmilor propusi in diverse cercetări e nevoie de baze de date care surprind modul de tastare care sa simuleze cat mai bine condițiile reale.

Scopul prezentei teze de doctorat este acela de a investiga cele mai mari platforme de MOOC, de a investiga amănunțit tematica keystroke dynamics in scopul autentificării continue a utilizatorilor sistemelor educationale, in special in timpul examenelor, de a furniza date empirice despre timpii de tastare colectate in condiții reale de la 80 de utilizatori, de a propune o structura de date eficienta ca memorie si ca performanta pentru reținerea datelor despre un utilizator si de a propune îmbunătățiri ale algoritmilor astfel încât sa crească performanta algoritmilor existenți.

Colectarea datelor despre ritmul de tastare al fiecărui utilizator s-a realizat pe web, cu autorul unui formular care a avut la baza algoritm scris in JavaScript. S-au colectat de la fiecare din cei 80 de utilizatori codul tastelor precum si cate doi timpi in milisecunde: timpul la care a fost apăsata fiecare tasta si timpul la care a fost ridicata fiecare tasta. Pe baza acestor informații s-a creat o baza de date, iar după prelucrare s-au obținut vectori de timpi care reprezentau fie timpul cat o tasta a fost apăsata, fie timpul scurs intre doua taste consecutive sau combinații ale acestor timpi. Pentru fiecare utilizator s-a făcut un pattern care reprezintă tiparul acestuia de a tasta anumite caractere folosite de regula cel mai frecvent, dar si di-graphs, perechi de cate doua caractere consecutive. In momentul simulării accesului unui utilizator la un anumit cont s-au calculat distantele dintre cei doi vectori de timp: patternul pentru contul respectiv si cel colectat de la utilizatorul care accesează contul. Performantele algoritmului dezvoltat in prezenta cercetare s-au calculat cu ajutorul Equal Error Rate (EER), indicator de performanta situat la intersecția False Acceptance Rate (FAR) si False Rejection Rate (FRR).

1.2 Obiectivele tezei

In prezentul proiect de cercetare, autorul a stabilit de la început patru obiective:

Obiectivul 1, O1, Primul obiectiv al prezentei teze este sa se colecteze o baza de date cu tiparul de testare de la cel puțin 80 de utilizatori, pentru a putea testa algoritmul de autentificare, dar si pentru a o pune la dispoziția altor cercetători interesați.

Obiectivul 2, O2, Al doilea obiectiv al prezentei teze este sa se implementeze un algoritm de autentificare a utilizatorilor unui calculator pe baza modului de tastare la tastatura.

Obiectivul 3, O3, Al treilea obiectiv al prezentei teze este sa propună cel puțin doua noi metrici de calcul a distantelor dintre doi vectori care sa genereze performante mai bune raportate la indicatorul de performanta Equal Error Rate (EER) decât metodele clasice.

Obiectivul 4, O4, Al patrulea obiectiv al prezentei teze este de a propune o structura de date cat mai eficienta ca spațiu, care sa conțină cele mai relevante informații despre modul de tastare al unui utilizator.

1.3 Structura tezei

Teza este organizată după cum urmează:

- Capitolul 1 prezintă contextul tezei, obiectivele tezei și structura tezei.
- Capitolul 2 prezintă state-of-the-art în domeniul căruia i se adresează această lucrare.

• Capitolul 3 prezintă metodologia de cercetare aplicată în acest proiect de cercetare. Etapele realizate în prezenta cercetare științifică sunt prezentate mai jos: A. Dezvoltarea platformei pentru achiziționarea datelor de intrare, B. Achiziționarea și prelucrarea inițială a datelor de intrare de la 80 de voluntari (modul de tastare la tastatură), C. Procesarea datelor colectate astfel încât să genereze un tipar de utilizator pentru fiecare utilizator, D. Dezvoltarea unui algoritm în limbajul de programare C pentru calcularea distanțelor utilizate în autentificarea dinamică pe baza modului de tastare, E. Simularea autentificării în sistem de către utilizatori autentici sau impostori pentru a măsura performanța algoritmului.

• Capitolul 4 este despre setul de date colectate pentru cercetarea de față. Acest capitol abordează validarea O1, formulat în primul capitol al acestei teze.

• Capitolul 5 În acest capitol este prezentat algoritmul de autentificare bazat pe modului de tastare liberă a textului la tastatură. În primul rând, este prezentată arhitectura algoritmului dezvoltat pentru prelucrarea datelor obținute de la utilizatori. Algoritmul simulează autentificarea utilizatorului pe baza modului de tastare și măsoară performanțele obținute. Dezvoltarea acestui algoritm este stabilită de O2, formulat în primul capitol al acestei teze.

• Capitolul 6 În acest capitol sunt prezentate o serie de experimente efectuate pentru a măsura performanța algoritmului dezvoltat în scopul acestei cercetări și pentru a analiza rezultatele obținute. Acest capitol abordează validarea O3 și O4, formulate în primul capitol al acestei teze.

• Capitolul 7 rezumă concluziile extrase din capitolele anterioare și direcțiile viitoare de cercetare în acest domeniu, pornind de la rezultatele prezentate în cuprinsul lucrării. Contribuțiile proprii ale autorului la domeniul autentificării pe baza modului de tastare sunt prezentate în acest capitol. Contribuția personală: propunerea a două metrici noi pentru calcularea distanței dintre doi vectori pentru a permite aproximarea gradului de asemănare între două tipare de la doi utilizatori diferiți sau de la același utilizator. De asemenea, datele colectate de la cei 80 de utilizatori despre modul de tastare de pe tastatură reprezintă o contribuție proprie, deoarece vor fi disponibile tuturor cercetătorilor interesați de efectuarea cercetărilor în domeniu. O altă contribuție proprie este propunerea unui tipar al utilizatorului pentru a stoca datele minime necesare pentru a obține performanțe în autentificarea continuă.

2. STATE-OF-THE-ART

2.1 Evoluția sistemelor educaționale

În acest subcapitol autorul prezintă evoluția platformelor MOOC (Massive Open Online Courses). În 2020, în platforma Coursera au fost înrolați 69 de milioane de cursanți [VAN20]. Numărul de cursuri online deschise masive a crescut în ultimii ani.

În acest capitol autorul face o introspecție în evoluția Massive Open Online Courses cu o comparație a celor mai importante platforme ale MOOC. De asemenea, în ultimii ani, cercetătorii au acordat atenție domeniului Learning Analytics [IVA16]. Avem din ce în ce mai multe date de la Learning Management Systems. Au existat provocări suplimentare vizibile în domeniul educației în 2020. Odată cu pandemia COVID-19, autoritățile nu au introdus doar restricții privind circulația cetățenilor, dar au înăsprit și măsurile care pun în aplicare noi reglementări cu referire la educație. Un număr mare de universități au trebuit să se adapteze la noile circumstanțe, mutând toate activitățile în mediul online. Aceste limitări au dus la un salt fără precedent în educația online. Dintr-o dată, atât profesorii, cât și studenții sau elevii, au fost nevoiți de condițiile nou implementate să-și mute întreaga activitate pe platforme educaționale online și să-și continue cursurile în acest mod. Acest proces a condus la dezvoltarea domeniului e-learning, ajutând la creșterea companiilor care sunt active în acest

domeniu și a obligat ca cei care nu au folosit aceste sisteme până acum să le învețe într-un ritm accelerat [IAP14a].

Sistemul educațional a evoluat continuu datorită inovațiilor tehnologice. În [DAN12] autorul a făcut o enumerare a inovațiilor: în 1841 tabla din clasă, în 1940 filmul, în 1957 televiziunea. Computerul a fost o altă invenție care a contribuit la evoluția educației. Internetul și tehnologiile de comunicare au dezvoltat, de asemenea, formatul educației [IAP14a].

Evoluția MOOC începe cu cursul „Connectivism and Connective Knowledge" - CCK08 în 2008, care a avut un număr mare, câteva mii de participanți online. Cursul a fost facilitat de Downes și Siemens [DOW14] [IAP14a]. Startul MOOC a fost în anul 2008, dar anul 2012 a adus o creștere mare, astfel încât a fost denumit anul MOOC. Anii următori după 2012 au fost ani buni pentru MOOC, cu milioane de cursanți și sute de parteneri implicați în dezvoltarea cursurilor [IAP14a]. Un număr record de utilizatori a apelat la învățarea online în 2020. Din martie, au existat peste 69 de milioane de înscrieri doar pe Coursera. O creștere de aproximativ 430% față de aceeași perioadă a anului trecut [VAN20] [IAP21b].

2.2 Analiza modului de tastare la tastatură – analiza literaturii în domeniu

Analiza modului de tastare (keystroke dynamics) este un domeniu de cercetare cu importanță din ce în ce mai mare în controlul accesului la rețea și securitatea cibernetică [ZHO12] [IAP21b]. Deocamdată, doar câteva studii se referă la analiza modului de tastare pentru textul tastat liber la tastatură, modul în care utilizatorii tastează ce text doresc. Majoritatea cercetărilor analizează doar un text predefinit, fix, text static [ZHO12] [SAL10] [ZAC10]. De regulă, textul predefinit, textul fix, poate fii numele de utilizator și parola, pentru că de fiecare dată se tastează același text [MON02]. Textul liber necesită două faze: faza de înrolare a utilizatorului în sistem și faza de verificare a utilizatorului [MON02].
În primul rând, utilizarea modului de tastare pentru identificarea utilizatorilor a fost cercetată în anii 1970 [ZHO12]. Spillane și-a scris concluziile despre prima cercetare din 1975 [FOR77] și Forsen, Nelson și Staron în 1977 [SPI75]. În cel de-al doilea război mondial, ritmul de tastare la telegraf al mesajului în codul morse a fost utilizat pentru a identifica dacă expeditorul mesajului este din partea aliaților sau din partea inamicilor. [BAN12] [VAC07] [DUN08] [IAP21a].

Analiza modului de tastare a fost studiată mai ales în legătură cu autentificarea, dar unele studii, cum ar fi [MES11], au studiat și detectarea stărilor emoționale ale utilizatorului care folosește tastatura. Alte studii se concentrează pe prezicerea vârstei și genului utilizatorilor, pe baza modului de utilizare a tastaturii și a mouse-ului [AVA17]. În [SAL18], autorii au explorat relevanța tiparelor de interacțiune individuale și generale ale tastaturii și mouse-ului și au modelat principii pentru a detecta emoția utilizatorilor în scenarii de învățare din lumea reală [IAP21a] . În [LIM14], autorii indică faptul că analiza automată a stresului cursantului din datele provenite de la mouse și tastatură este utilă pentru asigurarea adaptării în sistemele de e-learning [IAP21a].

Dacă majoritatea studiilor utilizează doar date preluate de la tastatură, există studii care utilizează o metodă mixtă de identificare a utilizatorului, bazată pe datele preluate de la tastatură, dar și pe datele preluate de la mouse [LOZ17]. Funcții suplimentare, cum ar fi presiunea apăsării tastelor, sunt utilizate în plus față de funcțiile bazate pe timp, dar pentru a captura aceste date este nevoie de ecrane tactile sau alte dispozitive speciale [TEH13]. Etapele unei cercetări în acest domeniu sunt, de regulă: extragerea caracteristicilor modului de tastare, crearea profilurilor utilizatorilor/ actualizarea acestora și identificarea criteriilor de eficiență ale algoritmilor [KOC19] [IAP21a].

Există și produse comerciale care se bazează pe analiza modului de tastare la tastatură.

În 2003, lucrarea [ILO03] prezintă compania BioNet Systems care a brevetat sistemul de autentificare BioPassword [ZIL98]. În România, Typing DNA este o companie, un start-up, care a primit fonduri de 6,2 milioane de euro în 2020 pentru a crea o identitate pe baza modului de a tasta pentru securitate [STE20].

Algoritmii de autentificare pe baza modului de tastare pot fi împărțiți în trei grupe majore: estimarea distanțelor, metodele statistice și machine learning. Metodele de recunoaștere a utilizatorilor pe baza modului de tastare utilizate în literatură sunt: bazate pe distanțele dintre utilizatori, rețele neuronale, statistice, probabiliste, machine learning, arborele decizional, calcul evolutiv, logică fuzzy sau altele [KOC19] [IAP21a].

Autorii de la [YUE04] concluzionează că utilizarea analizei modului de tastare poate face un sistem mai sigur.

3. METODOLOGIA CERCETĂRII

Etapele efectuate în prezenta cercetare științifică sunt descrise mai jos:

A. Dezvoltarea platformei pentru achiziționarea datelor de intrare,

B. Achiziționarea și prelucrarea inițială a datelor de intrare de la 80 de voluntari (modul de tastare la tastatură),

C. Procesarea datelor colectate astfel încât să genereze un tipar de utilizator pentru fiecare utilizator,

D. Dezvoltarea unui algoritm în limbajul de programare C pentru calcularea distanțelor utilizate în autentificarea dinamică pe baza modului de tastare,

E. Simularea autentificării în sistem de către utilizatori autentici sau impostori pentru a măsura performanța algoritmului

Primii doi pași ai metodologiei de cercetare, A. Dezvoltarea platformei pentru achiziționarea datelor de intrare și B. Achiziționarea și prelucrarea inițială a datelor de intrare de la 80 de voluntari (modul de tastare la tastatură), au rolul de a aborda O1, formulat în primul capitol al tezei: colectarea unei baze de date cu modelul de testare de la cel puțin 80 de utilizatori, pentru a testa algoritmul de autentificare pentru această cercetare, dar și pentru a-l pune la dispoziția altor cercetători interesați.

Al treilea pas al metodologiei de cercetare, C. Prelucrarea datelor de intrare astfel încât să genereze un model de utilizator pentru fiecare utilizator și ultimul pas al metodologiei de cercetare, E. Simularea autentificării sistemului de către utilizatori reali sau impostori pentru a măsura performanța algoritmul dezvoltat, are rolul de a aborda O4, așa cum este descris în primul capitol al tezei: să propună o structură de date cât mai eficientă, care să conțină cele mai relevante informații despre tiparul unui utilizator.

Al treilea pas al metodologiei de cercetare, C. Procesarea datelor colectate astfel încât să genereze un tipar de utilizator pentru fiecare utilizator, și al patrulea pas al metodologiei de cercetare, D. Dezvoltarea unui algoritm în limbajul de programare C pentru calcularea distanțelor utilizate în autentificarea dinamică pe baza modului de tastare, au rolul de a aborda O2, așa cum este descris în primul capitol al tezei: să implementeze un algoritm pentru autentificarea utilizatorilor unui computer pe baza modului de tastare.

Ultimul pas al metodologiei de cercetare, E. Simularea autentificării în sistem de către utilizatori autentici sau impostori pentru a măsura performanța algoritmului, are rolul de a aborda O3, așa cum este descris în primul capitol al tezei: să propună cel puțin două metrici noi pentru calcularea distanțelor dintre doi vectori care generează performanțe mai bune comparativ cu indicatorul de performanță Equal Error Rate (EER) decât metodele clasice.

4. SETUL DE DATE COLECTAT CU PRIVIRE LA MODUL DE TASTARE ÎN MOD LIBER CU SCOPUL AUTENTIFICĂRII

Pentru a cerceta analiza modului de tastare e nevoie de date de intrare obținute de la utilizatori ai calculatorului in diferite situații. Datele necesare sunt reprezentate de tastele tastate la tastatura dar si de timpii la care aceștia sunt apăsate. Timpul când o anumita tasta este apăsata, respectiv timpul la care o anumita tasta este ridicata. Diferența dintre acești timpi reprezintă timpul de apăsare al tastei. O alta informație importanta este timpul dintre doua taste. Diferența dintre timpul la care o tasta a fost lăsata libera si timpul la care se apasă următoarea tasta.

Aceste informații se pot obține doar într-un mediu controlat, cu acordul celor care participa la acest experiment. E nevoie de acord pentru ca se poate forma textul inițial pe care l-a scris utilizatorul de la tastatura având acces la aceste date, iar daca, de exemplu, este monitorizat un utilizator in timp ce trimite e-mail-uri sau face alte activități, s-ar putea ca informațiile sa fie confidențiale si sa nu vrea sa fie făcute publice.

In literatura de specialitate exista câteva seturi de date care sunt accesibile in scopul cercetării. Am apelat in prima faza la aceste seturi de date. Cele mai multe sunt reprezentate de texte in limba engleza, preluate din interiorul mediului universitar, de către cercetători de la colegii lor din universitate sau de la studenți. Unele seturi de date sunt preluate de un anumit program, într-un mediu special făcut pentru a prelua aceste date. Altele sunt făcute sa monitorizeze tot ce se tasteaza la un calculator, indiferent de programul utilizat la un moment dat de utilizator. Acesta monitorizează tot ce se tastează la tastatura si timpii de tastare indiferent daca utilizatorul scrie e-mail-uri, scrie într-un document Word, Excel sau programează la calculator într-un anumit mediu de programare.

In prezenta teza voi prezenta rezultatele analizării seturilor de date obținute de la lucrări in acest domeniu.

Pe de alta parte, in scopul cercetării am realizat propriul mediu de a obține date de la voluntari. Am creat un mediu web de preluare a tastelor si timpilor de tastare in JavaScript. Este creat un formular prin care se preiau tastele si timpii de tastare din timpul completării unui formular pe o pagina web. Pagina web a fost realizata pe platforma sites.google.com. Platforma web poate fi accesata la adresa https://sites.google.com/view/cataliniapa.

Pentru a capta tastele si timpii de tastare am creat un formular web prin care utilizatorii au fost invitați sa răspundă pe rând la mai multe întrebări generice. Textul introdus de la tastatura de fiecare utilizator este un text scris liber de fiecare utilizator in parte, fără a fi nevoie sa reproducă un anumit text predefinit. La fiecare text box au fost formulate o serie de întrebări generice care sa ghideze utilizatorul spre o anumita tema in textul pe care l-a completat. Întrebările formulate au fost despre vreme, despre ziua ideala sau despre sistemul de învățământ. Pentru a forma baza de date pentru cercetare nu este relevanta tema textului, ci modul in care acesta este scris.

Formularul creat pentru a achiziționa seturi de date in scopul cercetării a fost completat de un număr de 80 de utilizatori. Aceștia au furnizat date pentru 410.633 de evenimente ale tastelor. Media pe utilizator este de 5132 evenimente ale tastelor. Timpul total folosit de toți cei 80 de utilizatori pentru a completa formularul a fost de 23 de ore, 28 de minute si 19 secunde. Media timpului petrecut de utilizatori pe platforma de colectare a datelor este de 17 minute si 36 de secunde.

In programul scris pentru a face analiza datelor, pentru fiecare eveniment al tastei s-a folosit o structura pentru a retine cele 3 informații colectate. Cele trei informații sunt colectate ca si numere întregi.

Prima informație colectata despre un key event este key code. Fiecare tasta este codificata printr-un număr întreg cuprins intre 8 si 222, conform corespondentelor prezentate

in Table 3.

A doua informație colectata despre un key event este un număr întreg reprezentând evenimentul surprins. Evenimentul surprins poate fi Key Down sau Key Up. Evenimentul Key Down l-am codificat prin cifra 0, iar evenimentul Key Up l-am codificat prin cifra 1.

A treia informație colectata despre un key event este timpul la care s-a petrecut evenimentul de Key Down sau de Key Up. Preia timpul sistemului de calcul exprimat in milisecunde.

Din informațiile celor 410.633 de evenimente ale tastelor s-au refăcut caracterele tastate de către utilizatori de la tastatura. In total s-au tastat la tastatura un număr total de 200.299 de taste.

5. DEZVOLTAREA ALGORITMULUI PENTRU AUTENTIFICARE BAZATĂ PE MODUL DE TASTARE LA TASTATURĂ

Arhitectura sistemului de autentificare bazat pe modul de tastare al tastelor are doua parți importante: prima este partea de training a sistemului, parte in care utilizatorii se înrolează in sistem furnizând date despre modul cum tastează. In aceasta faza se creează cate un pattern pentru fiecare utilizator si e stocat in baza de date pentru a fi utilizate in faza de autentificare continua. Cea de-a doua este faza de autentificare continua. In aceasta faza sistemul verifica in mod continuu utilizatorii conectați cu un username si o parola valide. Pe tot parcursul timpului cat un user este conectat in cont, sistemul preia de la acesta date despre modul de tastare si compara in mod continuu patternul rezultat cu patternul din baza de date. Atat timp cat exista similitudine acceptabila intre cele doua pattern-uri utilizatorul rămâne logat in sistem. In momentul in care sistemul constata ca nu se mai aseamănă cele doua pattern-uri, cel preluat de la utilizatorul logat in cont si cel din baza de date, sistemul generează un semnal de alarma si utilizatorul este scos din cont. Acesta poate reintra in cont reintroducând username-ul si parola.

Colectarea si prelucrarea inițiala a datelor de intrare sunt primii pași realizați in scopul de a se obține key events ale fiecărui utilizator. Aceste date reprezintă datele de intrare pentru algoritmul de autentificare continua pe baza keystroke dynamics. După ce se stabilește dimensiunea semple size, pasul următor este de a se împărții datele de le utilizatori in secvențe de key events. Algoritmul transforma key events in informații despre taste si informații despre di-graph-uri, iar mai apoi formează vectorii de timp necesari pentru calculul distantelor. După ce au fost parcași pașii descriși mai sus se trece la calculul distantelor dintre vectori, pentru a se stabili similaritatea dintre doi utilizatori. Se folosesc 4 tipuri de distante: Euclidian distance, Manhattan distance, R distance si A distance. Cu aceste distante calculate pentru fiecare utilizator din baza de date se trece la simularea autentificării in sistem, pe rând de fiecare utilizator din baza de date. In urma simulării autentificării in sistem se generează 4 indicatori de performanta a algoritmului: False Acceptance Rate (FAR), False Rejection Rate (FRR), True Acceptance Rate (TAR) si True Rejection Rate (TRR). Pe baza acestora se poate calcula Equal Error Rate (EER), principalul indicator al performantei algoritmilor folosit in aceasta teza. De asemenea, pentru vizualizarea performantelor se generează doua grafice: graficul FAR și FRR și curba ROC.

6. EXPERIMENTE ȘI REZULTATE - SIMULAREA AUTENTICĂRII ÎN SISTEM DE CĂTRE UTILIZATORI AUTENTICI ȘI IMPOSTORI

În acest capitol sunt prezentate o serie de experimente efectuate pentru a măsura performanța algoritmului scris în scopul acestei cercetări și pentru a analiza rezultatele obținute. Treptat, sunt prezentate experimente cu timpul de apăsare a unei singure taste, în

subcapitolul 6.1, și experimente cu modul de apăsare a unei perechi de taste consecutive, în subcapitolul 6.2. Atât în analiza caracteristicilor cu o singură tastă, cât și cu perechi de taste consecutive, se calculează Equal Error Rate (EER) pentru a aprecia performanțele algoritmilor. Rezultatele sunt prezentate în cazul experimentelor folosind distanța euclidiană (în subcapitolele 6.1.1 și 6.2.3), distanța Manhattan (în subcapitolele 6.1.2 și 6.2.4), distanța R (în subcapitolul 6.1.3) și distanța A (în subcapitolele 6.1.4 și 6.2.5). Capitolul investighează, de asemenea, în subcapitolul 6.1.5 diferențele de performanță când pattern-ul utilizatorului este construit cu diferite dimensiuni ale secvenței de taste colectate, începând de la 200 evenimente ale tastelor / pattern și până la 3000 evenimente ale tastelor / pattern. La finalul capitolului, în urma tuturor experimentelor efectuate și prezentate, autorul propune, în subcapitolul 6.4, Propunerea de noi metrici pentru calcularea distanțelor dintre utilizatori, modificarea a două metrici obținând metrici noi pentru calcularea distanțelor dintre doi vectori cu care se obțin performanțe mai mari decât cu metodele clasice de calcul. Pentru cele două noi metrici, sunt prezentate performanțele obținute în termeni de Equal Error Rate (EER). Prin propunerea acestor noi metrici, O3 este validat. De asemenea, propune, în subcapitolul 6.5, Tiparul de utilizator propus, o structură pentru păstrarea pattern-ului unui utilizator, o structură care ocupă o memorie mică și necesită puțin timp pentru a efectua calculele necesare în algoritmi. Prin propunerea modelului de utilizator, O4 este validat. La sfârșitul capitolului, în subcapitolul 6.6, performanțele obținute în prezenta cercetare sunt comparate cu cele obținute de alți autori în cercetările lor.

7. CONCLUZII ȘI DIRECȚII DE CERCETARE VIITOARE

Prezenta cercetare si-a propus aprofundarea domeniului de autentificare cu ajutorul modului de a tasta liber un text la tastatură, in special pentru platforme de educație online. De asemenea, la începutul cercetării s-au formulat 4 obiective care au fost atinse rând pe rând pe parcursul cercetării.

In capitolul 4 au fost prezentate aspectele referitoare la platforma creata pentru a colecta date de la cei 80 de utilizatori. S-a prezentat platforma web de colectare a datelor si codul acesteia scris in JavaScript, precum si modul de colectare si transfer al acestor date. Prin crearea bazei de date cu datele despre modul de tastare a 80 de utilizatori si prin punerea acesteia la dispoziția altor cercetări in domeniu s-a abordat O1, formulat in primul capitol al tezei.

La începutul capitolului 5 s-a prezentat algoritmul realizat pentru a simula autentificarea utilizatorilor pe baza modului de tastare la tastatură. Datele de intrare utilizate ca baza pentru algoritm au fost colectate prin platforma prezentata in cadrul capitolului anterior, capitolul 4. Prin dezvoltarea algoritmului prezentat O2 a fost abordat.

In cadrul subcapitolului 5.5 Proposing new metrics for calculating distances between users, s-au prezentat cele doua propuneri de metrici modificate care genereaza rezultate mai bune, ca si performanta a indicatorului Equal Error Rate (EER). O3 a fost abordat in cadrul acestui subcapitol. Cu ajutorul celor 2 metrici se îmbunătățesc performantele algoritmului de autentificare cu 25,24%, respectiv cu 37,47%. Cea mai buna performanta obținuta cu metrica modificata este de EER=3,27%.

In cadrul subcapitolului 5.6, Proposed user pattern, s-a prezentat propunerea de pattern al utilizatorului rezultata in urma experimentelor si a performantelor obținute, astfel încât informația care retine caracteristicile unui utilizator sa ocupe spațiul de memorie optim si sa poată contribui la un algoritm rapid. Se propune o structura care retine media si deviație standard, a timpului de apăsare a tastei, pentru cele mai frecvent folosite 14 litere si media timpilor de tastare a primei taste, a celei de-a doua taste si a timpului total pentru cele mai frecvent folosite 12 di-graph-uri (perechi de taste consecutive). Structura astfel obținuta ocupa

256 octeți în memorie pentru fiecare utilizator. Aceasta propunere a abordat O4.

7.1.1 Contribuțiile personale prezentate în această cercetare sunt:

1. A fost dezvoltat un algoritm de autentificare continuă pe baza modului de tastare la tastatură. Algoritmul poate fi găsit în Anexa 1 - Algoritm de dinamică a tastării textului liber pentru autentificare continuă și a fost prezentat în Capitolul 4.

2. A fost creată o bază de date cu modul de tastare de la 80 de utilizatori, au fost colectate în total 410.000 evenimente ale tastelor într-un timp total de aproximativ 24 de ore. Detaliat în capitolul 5

3. A fost propusă o metrică modificată pornind de la distanța clasică Manhattan, calculată pe cele mai utilizate 14 litere. Noua metrică propusă îmbunătățește coeficientul de performanță EER de la 7,13% la valoarea de 5,33%. Aceasta înseamnă o îmbunătățire a performanței cu 25,24%. Detalii despre metrica propusă sunt în subcapitolul 6.5 Propunerea unor noi metrici pentru calcularea distanțelor între utilizatori, 6.4.1 Nouă metrică pentru calcularea distanțelor pe baza timpului unei taste individuale.

4. A fost propusă o metrică modificată pentru calculul distanței, cu folosirea celor mai utilizate 12 perechi de taste consecutive (di-graph-uri). Noua metrică îmbunătățește coeficientul de performanță EER de la 5,23% la valoarea de 3,27%. Aceasta înseamnă o îmbunătățire a performanței cu 37,47%. Detalii despre noua valoare propusă se află în subcapitolul 6.4 Propunerea unor noi metrici pentru calcularea distanțelor dintre utilizatori, 6.4.2 Nouă metrică pentru calcularea distanțelor pe baza perechilor de taste consecutive (di-graph-uri).

5. A fost propusă o structură pentru pattern-ul utilizatorului cu eficientizarea spațiului utilizat, dar și cu premisele pentru a face calculele necesare într-un timp scurt. Spațiul total ocupat de un astfel de pattern pentru un utilizator este de doar 256 de octeți. Propunerea formulată pentru stocarea pattern-ului este reprezentată în subcapitolul 6.5 Pattern-ul utilizatorului.

7.2 Direcții de cercetare viitoare

Teză și-a atins obiectivele stabilite, dar noi teme de cercetare pot continua prezenta cercetare, precum:

- Extinderea bazei de date cu modul de tastare prin colectarea datelor de la un număr mai mare de utilizatori;
- Extinderea bazei de date prin colectarea datelor de la cei 80 de utilizatori în sesiuni noi pentru a cerceta evoluția tiparului de tastare în timp
- Analiza prin dezvoltarea de noi algoritmi, pe baza celorlalte tehnici de a calcula similaritatea dintre utilizatori
- Aplicarea metricilor propuse în această lucrare asupra datelor din alte baze de date disponibile din cercetări științifice
- Analiza particularităților literelor speciale din limba română, care nu se regăsesc în limba engleză: Ă, Î, Â, Ș, Ț.
- Analiza semnelor de punctuație sau a celorlalte taste diferite de litere, SPACE, ENTER, TAB, BACKSPACE etc.
- Modificarea condițiilor de colectare a datelor: schimbarea tastaturii, sub stres etc.
- Analiza perechii de taste consecutive (di-graph) în strânsă legătură cu cuvântul în care apare
- Dezvoltarea algoritmilor de autentificare pe baza modului de tastare, prin analiza timpilor generați de grupuri de trei taste consecutive (tri-graph-uri)
- Dezvoltarea algoritmilor de autentificare pe baza modului de tastare, prin analiza

timpilor generați de grupuri de n taste consecutive (n-graph-uri)

BIBLIOGRAFIE SELECTIVĂ

[ARW17] Arwa Alsultan, Kevin Warwick, Hong Wei, Non-conventional keystroke dynamics for user authentication, Pattern Recognition Letters, Volume 89, 2017, Pages 53-59, ISSN 0167-8655

[AVA17] Avar Pentel. 2017. Predicting Age and Gender by Keystroke Dynamics and Mouse Patterns. In Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17). Association for Computing Machinery, New York, NY, USA, 381–385. DOI:https://doi.org/10.1145/3099023.3099105

[BAN12] Banerjee, Salil & Woodard, D.L.. (2012). Biometric Authentication and Identification Using Keystroke Dynamics: A Survey. Journal of Pattern Recognition Research. 7. 116-139. 10.13176/11.427.

[DAN12] J. Daniel. 2012. Making Sense of MOOCs: Musings in a Maze of Myth, Paradox and Possibility. Technical Report. Korea National Open University. http://www.tonybates.ca/wp-content/uploads/Making-Sense-of-MOOCs.pdf Retrieved February 2014

[DOW14] Downes, S. 2008. Places to go: Connectivism & Connective Knowledge. Innovate 5 (1). http://www.innovateonline.info/index.php?view=article&id=668 January 2014

[DUN08] T. Dunstone and N. Yager. Biometric System and Data Analysis: Design, Evaluation, and Data Mining. Springer, 1 edition, 2008.

[FOR77] G. Forsen, M. Nelson, and R. Staron, Jr. "Personal attributes authentication techniques", Technical Report RADC-TR-77-333, Rome Air Development Center, October 1977.

[IAP14a] Iapa, A.C. (2014), Outstanding research in MOOC and future development, Proceedings of the 10th International Scientific Conference "eLearning and Software for Education" Bucharest, Editura Universitatii Nationale de Aparare "Carol I" 2014 Volume 1, 251-254, DOI: 10.12753/2066-026X-14-035

[IAP21a] Iapa A.C., Cretu V.I., Modified Distance Metric That Generates Better Performance For The Authentication Algorithm Based On Free-Text Keystroke Dynamics, IEEE 15th International Symposium on Applied Computational Intelligence and Informatics, Timisoara, Romania, 2021 – paper sent, unpublished

[IAP21b] Iapa A.C., Cretu V.I., Evaluating the performance of authentication algorithms based on keystroke dynamics used in online educational platforms, The 17th International Scientific Conference eLearning and Software for Education, Bucharest, Romania, 2021 – paper sent, unpublished

[ILO03] Ilonen, Jarmo. (2003). Keystroke dynamics. Advanced Topics in Information

processing–lecture (2003).

[IVA16] Ivanova, Malinka, Holotescu, C., Grosseck, G., Iapa, C. "RELATIONS BETWEEN LEARNING ANALYTICS AND DATA PRIVACY IN MOOCs." The International Scientific Conference eLearning and Software for Education. Vol. 3. " Carol I" National Defence University, 2016.

[KOC19] Kochegurova, Elena & Luneva, Elena & Gorokhova, Ekaterina. (2019). On Continuous User Authentication via Hidden Free-Text Based Monitoring: Volume 2. 10.1007/978-3-030-01821-4_8.

[LIM14] Y. M. Lim, A. Ayesh and M. Stacey, "Detecting cognitive stress from keyboard and mouse dynamics during mental arithmetic", Proc. Sci. Inf. Conf. (SAI), pp. 146-152, Aug. 2014.

[LOZ17] Lozhnikov, Pavel & Sulavko, Alexey & Ekaterina, Buraya & Viktor, Pisarenko. (2017). Authentication of Computer Users in Real-Time by Generating Bit Sequences Based on Keyboard Handwriting and Face Features. Voprosy kiberbezopasnosti. 24-34. 10.21681/2311-3456-2017-3-24-34.

[MES11] A. Messerman, T. Mustafic, S. A. Camtepe and S. Albayrak. Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics. Proceedings of IEEE International Joint Conference on Biometrics. 1–8, 2011

[MON02] Monrose F, Reiter MK, Wetzel S (2002) Password hardening based on keystroke dynamics. Int J Inf Secur 1(2):69–83

[ROT14] J. Roth, X. Liu and D. Metaxas, "On Continuous User Authentication via Typing Behavior," in IEEE Transactions on Image Processing, vol. 23, no. 10, pp. 4611-4624, Oct. 2014, doi: 10.1109/TIP.2014.2348802.

[SAL10] E. Al Solami, C. Boyd, A. Clark, and A. K. Islam, "Continuous Biometric Authentication: Can It Be More Practical?", IEEE Int'l Conf. on High Performance Computing and Communications (HPCC), pp. 647-652, 2010.

[SAL18] S. Salmeron-Majadas, R. S. Baker, O. C. Santos and J. G. Boticario, "A Machine Learning Approach to Leverage Individual Keyboard and Mouse Interaction Behavior From Multiple Users in Real-World Learning Scenarios," in IEEE Access, vol. 6, pp. 39154-39179, 2018, doi: 10.1109/ACCESS.2018.2854966.

[SPI75] R. Spillane, "Keyboard Apparatus for Personal Identification", IBM Technical Disclosure Bulletin, vol. 17, no. 3346, 1975.

[STE20] Stefan Koritar. (2020). Romanian startup Typing DNA raises €6.2 million in Series A funding to create ‘typing identity’ for security (2020).

[TEH13] Teh, Pin Shen & Teoh, Andrew & Yue, Shigang. (2013). A Survey of Keystroke Dynamics Biometrics. TheScientificWorldJournal. 2013. 408280. 10.1155/2013/408280.

[UMP85] D. Umphress and G. Williams, "Identity Verification through Keyboard Characteristics", Int'l J. Man-Machine Studies, Vol. 23, No. 3, pp. 263-273, 1985.

[VAC07] J. R. Vacca. Biometric Technologies and Verification Systems. Butterworth-Heinemann, 1 edition, 2007.

[VAN20] Vandenbosch, B., Most Popular Courses of 2020: A Year of Mental Health, Contract Tracing, and Job-Relevant Skills, Coursera Blog.

[YUE04] Yu, Enzhe & Cho, Sungzoon. (2004). Keystroke dynamics identity verification - Its problems and practical solutions. Computers & Security. 23. 428-440. 10.1016/j.cose.2004.02.004.

[ZAC10] R. Zack, C. Tappert, and S. Cha, "Performance of a long-text-input keystroke biometric authentication system using an improved k-nearest-neighbor classification method", IEEE Int'l Conf. on Biometrics: Theory Applications and Systems (BTAS), pp. 1-6, 2010.

[ZHO12] Y. Zhong, Y. Deng and A. K. Jain, "Keystroke dynamics for user authentication," 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, 2012, pp. 117-123, doi: 10.1109/CVPRW.2012.6239225.

[ZHO15] Zhong, Yu & Deng, Yunbin. (2015). A Survey on Keystroke Dynamics Biometrics: Approaches, Advances, and Evaluations. 10.15579/gcsr.vol2.ch1.

[ZIL98] Zilberman, A.G.: Security method and apparatus employing authentication by keystroke dynamics (1998) United States Patent 6,442,692.