(19) OFICIUL DE STAT
PENTRU INVENŢII ŞI MĂRCI
Bucureşti

ROMÂNIA

(11) **RO 135860 A2**
(51) Int.Cl.
G06F 21/60 (2013.01);
H04W 12/02 (2009.01)

(12) # CERERE DE BREVET DE INVENŢIE

(21) Nr. cerere: **a 2020 00800**

(22) Data de depozit: **02/12/2020**

(41) Data publicării cererii:
30/06/2022 BOPI nr. **6/2022**

(71) Solicitant:
• **REPSMATE SOFTWARE S.R.L.**,
BD.CHIŞINĂU, NR.18, BL.M8, SC.1, ET.10, AP.129, SECTOR 2, BUCUREŞTI, B, RO

(72) Inventatori:
• **ROŞCA ALIN-GABRIEL**,
ŞOS.ŞTEFAN CEL MARE, NR.30, BL.26, SC.1, ET.9, AP.26, SECTOR 2, BUCUREŞTI, B, RO

(74) Mandatar:
STRENC SOLUTIONS FOR INNOVATION S.R.L., *STR.LUJERULUI NR.6, BL.100, SC.B, ET.3, AP.56, SECTOR 6, BUCUREŞTI*

(54) # SISTEM ŞI METODĂ PENTRU ANONIMIZAREA DATELOR DE IDENTIFICARE A PERSOANELOR AFLATE ÎNTR-O CONVORBIRE AUDIO/VIDEO

(57) **Rezumat:**

Invenţia se referă la un sistem şi o metodă pentru anonimizarea datelor cu caracter personal ale persoanelor aflate într-o convorbire audio/video, care permit comunicarea între două sau mai multe entităţi fără a fi necesar să se solicite acordul de înregistrare a convorbirii, în timp ce datele cu caracter personal ale participanţilor sunt anonimizate ireversibil, iar restul conversaţiei poate să fie stocată şi folosită în alte scopuri, productive. Soluţia conform invenţiei are la bază o tipologie generală a unei reţele de comunicaţii (**RC**) care transmite şi primeşte informaţie printr-un protocol de comunicare (**PC**) care poate să faciliteze traficul de date, voce, video şi nu numai, denumite în continuare conversaţie, accesând diferite canale între diferite dispozitive şi aplicaţii de tip software, şi constă într-un subsistem de anonimizare a conversaţiilor (**SAC**) ce se conectează la traficul dintre reţeaua de comunicaţii (**RC**) şi protocolul de comunicare (**PC**) prin intermediul unei interfeţe de programare a aplicaţiei (**API**) şi are rolul de a prelua conversaţia şi de a anonimiza ireversibil datele cu caracter personal, ulterior anonimizării, conversaţia fiind preluată de un subsistem de transmitere (**SDT**) care o transmite, după caz, către un server de analiză (**SDA**) şi/sau către un server de stocare (**SDS**).

Revendicări: 5
Figuri: 4

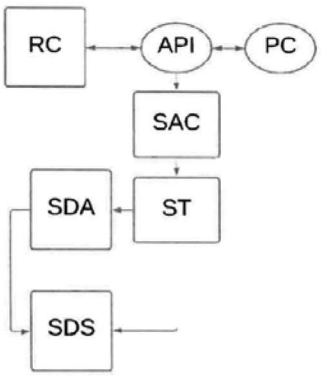

Fig. 1

RO 135860 A2

OFICIUL DE STAT PENTRU
Cerere de brevet de invenţie
Nr. a 2020 00800
Data depozit 02-12-2020

39 38

Sistem si metoda pentru anonimizarea datelor de identificare a persoanelor

aflate intr-o convorbire audio/video

Sistemul si metoda conform inventiei, permite comunicarea dintre doua sau mai multe entitati, de exemplu companie-client, fara sa fie necesar cererea acordului de inregistrare a convorbirii, din cauza ca datele cu caracter personal sunt anonimizate ireversibil, iar continutul conversatiei poate sa fie stocat si folosit in alte scopuri productive pentru companie.

Solutia tehnica care sta la baza inventiei se poate aplica in orice domeniu in care se regasesc conversatii audio si/sau video dintre doua sau mai multe persoane, atat in timp real cat si inregistrate si are in vedere un sistem de tip soft conectat la arhitectura de comunicare a unei entitati, companie sau instutie, arhitectura de comunicare care la randul ei poate integra diverse canale de comunicatie precum sistem de mesagerie audio sau text de tip SMS, dispozitive mobile sau calculatoare, conectate sau nu prin internet, sisteme de conferinte sau telefonie de tip video si/sau audio, sau dispozitive fixe de inregistrare.

Se cunosc sisteme de anonimizare a datelor cu caracter personal sau caracter sensibil, al utilizatorilor care isi dau acordul ca datele acestora sa fie prelucrate, insa este de mentionat ca practic, la ansamblul solutiilor existente obiectivul final al solutiei noastre este aproape imposibil de atins, de exemplu: intr-o discutie de vanzari la care scriptul si linia discutiei cu clientii sunt foarte bine structurate, cu o tinta foarte clara, iar prin solicitarea acordului clientului, structura mentionata anterior poate sa fie debalansata, iar tinta discutiei poate sa nu mai fie atinsa.

Astfel, solutia din documentul de brevet european EP 3 644 253 A1 "Anonymisation of identifier" are scopul de a anonimiza identificatori unici, prin generarea de identificatori unici anonimizati, care pot recupera ulterior identificatorul unic original, unde identificatorul unic este de fapt utilizatorul. Desi solutia propusa in document se refera la faptul ca datele utilizatorului sunt anonimizate, ea are dezavantajul ca aceste date pot fi aflate momentul in care se doreste acest lucru.

Este cunoscuta de asemenea, solutia din documentul de brevet european EP 3 651 157 " Processing of anonymised patient data generated by a mobile medical device " in care face referire la datele utilizatorilor anonimizate, care pot fi aduse la forma intiala respectiv deanonimizate, cu acordul utilizatorilor. In acest

caz utilizatorul isi da oricum acordul de a folosi dispozitivul medical, cu scopul de a oferi datele acestuia spre o analiza mai complexa.

Dezavantajul solutiilor mentionate este acela ca utilizatorul isi da acord ca datele lui sa fie folosite si in consecinta exista consecinta negativa ca in cazul unui atac cibernetic, atacatorii sa poata deanonimiza datele acestora.

Problema tehnica pe care o rezolva inventia este de a facilita utilizarea informatiilor provenite dintr-o conversatie monitorizata in timp real, cu scopul de imbunatatire a activitatilor companiilor, fara a cere acordul participantilor in discutie ca respectiva conversatie sa fie inregistrata si implicit cu respectarea integrala a prevederilor din Regulamentul General Privind Protectia Datelor GDPR.

Sistemul si metoda de anonimizare conform inventiei, are rolul de a transforma multe procese care pot fi realizate manual, la un cost foarte mare, intr-un proces automat care poate fi realizat eficient la un cost mic si permite organizatiilor sa aiba incredere in securitatea datelor lor si sa respecte protocoalele si reglementarile impuse de Regulamentul General Privind Protectia Datelor GDPR.

Date cu caracter personal, conform Regulamentul General Privind Protectia Datelor GDPR inseamna orice informatii privind o persoana fizica identificata sau identificabila ("persoana vizata"); o persoana fizica identificabila este o persoana care poate fi identificata, direct sau indirect, in special prin referire la un element de identificare, cum ar fi un nume, un numar de identificare, date de localizare, un identificator online, sau la unul sau mai multe elemente specifice, proprii identitatii sale fizice, fiziologice, genetice, psihice, economice, culturale sau sociale;

Conform Regulamentul General privind Protectia Datelor GDPR datele sunt anonimizate, atunci cand orice element de identificare dispare. Daca un element ramas poate servi, singur sau impreuna cu altele, la reidentificarea persoanei, atunci anonimizarea nu a fost facuta, iar datele nu ar trebui pastrate. Odata ce datele au fost anonimizate complet, acestea ies din sfera de protecție a Regulamentului.

Solutia tehnica, se bazeaza o structura functionala care permite inregistrarea si transcrierea discutiilor provenite din arhitectura de comunicatii si anonimizarea ireversibila, in timp real, a datele cu caracter personal si deci fara ca datele utilizatorilor sau a participantilor sa mai poata fi identificabile, sau implicit sa fie cerut acordul acestora pentru inregistrarea discutiilor, respectiv indeplinirea unei caracteristici tehnice functionala care nu este posibil a fi realizata in baza solutiilor cunoscute.

Functionarea sistemului si implicit derularea metodei de anonimizare, se deruleaza prin conectarea sistemului soft la arhitectura de comunicatie, interceptarea conversatiei dintre doua sau mai multe persoane, inregistrarea discutiei temporar si transcrierea a datelor in timp real, detectarea datelor cu caracter personal, urmata de anonimizarea ireversibila a acestora si repectiv transmiterea acestora catre serverele dedicate de stocare sau analiza.

Selectia datelor care trebuiesc anonimizate se realizeaza printr-o arhitectura conectata ponderata paralel, in care ponderile pot fi setate in functie de cazul utilizarii. Fiecare modul din aceasta arhitectura calculeaza probabilitatea ca un token sa fie sau nu informatie care trebuie anonimizata, iar suma ponderata a probabilitatilor returneaza un potential global care comparat cu un trashold, setat in functie de cazul utilizarii aplicatiei, conduce la decizia de a considera un token sensibil sau nu. Ponderile sunt necesare pentru a modula importanta componentelor si aceasta arhictectura acopera si cazurile in care unul sau doua din module lipsesc, echivalent cu ponderea zero.

Solutia tehnica conform inventiei, prezinta urmatoarele avantaje:

- Imbunatatirea comunicarii sau proceselor companiilor pentru a creste satisfactia clientilor sau persoanelor cu care interactioneaza;
- Evolutie tehnologica prin antrenarea de modele de inteligenta artificiala care sa raporteze informatiile analizate si sa asiste reprezentantii companiilor;
- Securitate garantata a informatiilor stocate, in concordanta cu Regulamentul General privind Protectia Datelor GDPR, fara a exista riscul de furt sau impartasire a informatiilor, persoanelor neautorizate.

Se da in continuare un exemplu de realizare a inventiei, in legatura si cu Fig.1...Fig. 4 care reprezinta:

Fig.1 - Arhitectura de ansamblu a Sistemului de Anonimizare a Conversatiilor – ASAC;

Fig.2 - Interfata Sistemului de Anonimizare a Conversatiilor cu Participanti – ISACP;

Fig.3 - Subsistemul de anonimizare a conversatiilor – SAC;

Fig. 4 - Modulul de Detectare al Datelor din Conversatie – MDDC.

Fig. 1 conform inventiei ilustreaza o tipologie generala a unei Retele de Comunicatii (RC) care transmite si primeste informatie printr-un Protocol de Comunicare (PC). Acesta din urma poate sa faciliteze traficul de date, voce, video si nu numai, denumite in continuare conversatie, accesand diferite canale intre diferite

dispozitive si aplicatii de tip software. Inventia se conecteaza la traficul dintre Reteaua de Comunicatii (RC) si Protocolul de Comunicare (PC) printr-o Interfata de Programare a Aplicatiei (API) destinata interceptarii aceste conversatii si oferita de compania care faciliteaza Reteaua de Comunicare (RC) contractata de catre Utilizator (U).

Solutia tehnica conform inventiei, poate fi integrata cu aceasta interfata, dar nu limitata de aceasta, printr-un Subsistem de Anonimizare al Conversatiilor (SAC) care preia aceste conversatii si anonimizeaza ireversibil datelor cu caracter personal. Ulterior anonimizarii, conversatiile sunt preluale de un Subsistem de Transmitere (SDT) care le trimite, dupa caz, catre un Server de Analiza (SDA) si/sau un Server de Stocare (SDS). Atat Serverul de Analiza (SDA) cat si Serverul de Stocare (SDS) pot apartine unei alte companii contractate sau subcontractate de catre utilizator. Aceste conversatii pot fi transmise atat in timp real, in timpul conversatiilor, cat si dupa ce conexiunea a fost terminata.

Fig. 2 denumita Interfata Sistemului de Anonimizare al Conversatiilor cu Participanti (ISACP), ilustreaza persoanele care realizeaza traficul de conversatii. Se definesc ca si Utilizatori (U), orice entitate, fie companie, institutie sau individ si nu numai, care folosesc Arhitectura Sistemului de Anonimizare al Conversatiilor (ASAC), pentru a anonimiza datele sensibile sau al altor Participanti (P) care sunt prezenti in conversatii.

Se definesc ca Participanti (P) orice entitate, fie companie, institutie sau individ, dar nu limitati de acestia sau numarul lor, care trebuie sa isi dea acordul ca aceaste conversatii sa fie capturate, inregistrarate, structurate, stocate si nu numai, iar fara acordul acestora legea de protectie a datelor cu caracter personal fiind incalcata.

Conform Subsistemului de Anonimizare al Conversatiilor, detaliat in Fig. 3 functia Start reprezinta activarea Subsistemului de Anonimizare al Conversatiilor (SAC), realizat automat in momentul in care Protocolului de Comunicare (PC) primeste un trafic de pachete. Un Modul de Captare al Traficului (MCT) intercepteaza si capteaza conversatia automat, fara a fi necesara interventia altui modul. Un Modul de Inregistrare Temporara al Conversatiei (MTTC) are scopul de a pastra datele originale pana acestea sunt identificate si anonimizate. Acesta este precedat functional de un Modul de Transcriere in Date (MTD), care poate transcrie conversatiile de exemplu in date audio (reprezentate prin unde, valuri, date meta sau alte date caracteristice formatului audio), date text (reprezenat de un transcript scris in cuvinte), dar nu limitate de acestea. Ulterior procesului de transcriere este activat un Modul de Detectare al Datelor din Conversatie (MDDC) care urmareste sa gaseasca date cu caracter personal, dar nu limitat doar de acestea, care pot

duce la identificarea participantilor din cadrul conversatiei si are rolul de a asigura ca toate datele acestora sunt anonimizate ireversibil, conform legislatiei in vigoare care se refera la Regulamentul de Protectie al Datelor Personale GDPR, dar nu limitat de catre acesta. Fiind urmarita pastrarea doar a datelor din conversatie referitoare la compartament, abordare, dar nu limitate de aceste; facand parte cu proponderenta din categoria de comercializare a produselor si serviciilor. Colectarea acestor date au scopul de a instrui reprezentantii utilizatorilor si de a creste satisfactia participantilor din conversate. Astfel nemaiexistand procesul de prelucrare al datelor cu caracter personal. Dupa ce aceste date, ale Participantilor (P), au fost detectate, acestea trec printr-un Submodul de Anonimizare Ireversibila al Datelor (SAID), acesta folosind procedeuri diferite pentru tipuri diferite de date, astfel incat datele originale sa nu mai poata fi acesate si implicit sa nu poata duce la identificarea participantilor din discutie. Submodulul de anonimizare ireversibila a datelor (SAID) poate lucra, dar nu este limitat de a lucra, impreuna cu alte submodule la stergerea datelor, astfel incat acestea pot fi folosite separat in functie de nevoile utilizatorilor. In momentul in care datele au fost anonimizate, un Modul de Stergere al Datelor Temporare (MSDT) este activat si informatiile originale sunt sterse. Un Modul de Transmitere (MT) are rolul de a transmite in timpul discutie sau la terminarea acesteia, toate datele colectate, dupa caz, catre un Modul de Analiza din Serverul de Analiza (MASDA) si/sau un Modul de Stocare din Serverul de Stocare (MSSDS). Toata aceasta procedura functioneaza fara a se opri pana in momentul in care este activat un Modul Incheiere Discutie (MID) care detecteaza automat cand Protocol de Comunicare (PC) nu mai primeste trafic de pachete, astfel rezultand ca discutia a fost incheiata si este activata functia Stop.

Conform Modulului de Detectare al Datelor din Conversatie (MDDC), fig. 4 functia Intrare Date, reprezinta datele provenite din Modulul de Transcriere in Date (MTD) care trebuiesc analizate si avand ca obiectiv principal recunoasterea si clasificarea datelor cu caracter personal, avand in vedere clasa acestora (nume, prenume, organizatie, oras, tara, numar personal de identificare, informații socio-economice etc.). Acest lucru se realizeaza incepand cu un Submodul de Preprocesarea (SP1), responsabil de pregatirea datelor pentru urmatorul submodul si este una dintre cele mai importante sarcini ale procesarii limbajului natural (NLP). Acest Submodul de Preprocesare (SP1) este compus dintr-un Proces de Segmentare (PS) cu rolul de segmentare a fiecarei propozitii pentru ca aceasta sa fie procesata individual, fara sa depinda de contextul general al conversatiei.

Acesta interactioneaza cu un Proces de Tokenizare (PT) care imparte datele prin parametrizare in unigrame, de exemplu pentru datele transcrise in text, datele sunt impartite in cuvinte individuale sau seturi

de cuvinte. In cazul inventiei noastre tokenizarea cuvintelor individuale este reprezentata de spatiul regasit dupa fiecare cuvant transcris individual, iar seturile de cuvinte sunt reprezentate prin schimbarea vorbitorului. Procesul mentionat este urmat de un Proces de Analiza Morfosintactica (PAM) care analizeaza unigramele, pentru a le clasifica si eticheta la nivel semantic pentru a oferi o structura a accestor etichete sub forma clasificata. In functie de cazul de utilizare (client, set de date samd.), urmatoarele trei blocuri calculeaza probabilitatea ca un token sa fie sau nu informatie personala, care trebuie anonimizata. Ponderile sunt necesare pentru a modula importanta celor trei componente. Spre exemplu, e posibil ca pentru anume aplicatii, Modelul Bazat pe Reguli sa mearga mai bine, deci ponderea lui va fi mai mare, pentru altele mai greu de formalizat, ori Modelul Bazat pe Invatare Automata o sa aiba pondere mai mare.

Primul bloc este reprezentat de un Model Bazat pe Reguli (MBR) care implementeaza diferite reguli pentru a descoperi unele clase de entitati, care pot fi asociate cu date sensibile legate de categoria numere (ex: cod numeric personal, numar de telefon, numar de asigurare, cod unic de identificare, cod postal etc.), adrese de e-mail și unele formate de data (ex: data de nastere). In plus fata de aceste reguli exista in unele cazuri o validare suplimentara. Aceasta validare se efectueaza pe toate numerele personale in care exista o validare de control, cifra de verificare sau suma de verificare. Aceasta ne permite sa dezambiguizam si sa avem o mai mare certitudine a cazurilor, cum ar fi numarul de telefon (ex : 0040712345678) și codul numeric personal (ex : 1234567890123), ambele continand 13 cifre. Pentru numarul de telefon, de exemplu, au fost create un set de reguli și un set de cuvinte de context. Regexul (o expresie regulata ; este o secventa de caractere care definesc un model) utilizat pentru extragerea acestei entitati pentru Romania a fost : 13 numere sub forma: prefix [0040] cu sinonim [+40], urmat de {9} cifre din categoria [0-9]. In timp ce lista de cuvinte contextuale pentru aceasta poate fi, dar nu limitata de aceasta : numar de telefon, numar de contact, telefon, telefon mobil, fax, telefon fix, contact. Lista de cuvinte contextuale se compun dintr-un set de cuvinte care apar de obicei in contextul unui numar de telefon. Aceasta a fost adaugata modelului pentru a rezolva erorile din unele tipuri de date deoarece consta dintr-un cuvant sau lista de cuvinte specifice pentru fiecare clasa de entitate care trebuie sa existe in text pentru a confirma rezultatul obtinut.

Cel de-al doi-lea bloc este un Model Bazat pe Lexicon (MBL) care combina rezultatele Procesului de Analiza Morfosintactica (PAM), un set de lexicoane, tehnici de derivatie si lematizare. Scopul fiind recunoasterea claselor de entitati (date cu caracter personal dar nu limitate de acestea). Pentru fiecare entitate, folosim lexicoane diferite cu propriile lor caracteristici specifice. Acest tip de implementare consta in compararea tokenilor prezenti in text cu lexiconul si intelegerea daca acestea corespund aceleiasi

entitati. Urmari ca toate entitatile sa fie implementate cu modele de invatare automata, cu scopul de a intelege cum se pot obtine cele mai bune rezultate pentru fiecare clasa de date.

Cel de-al trei-lea bloc este un Model Bazat pe Invatare Automata (MBIA) folosit pentru descoperirea claselor de entitati cu o ambiguitate mai ridicata sau pentru cazurile in care nu exista reguli bine definite. Acest ultim model avand cele mai bune rezultate datorita retelelor neuronale. Folosind aceasta arhictectura putem sa acoperim si cazurile in care unul sau doua din cele trei module lipsesc, echivalent cu ponderea zero. Ulterior analizei celor trei blocuri se activeaza un Submodul Bloc de Functii care calculeaza suma ponderata a probabilitatilor pentru a detecta potentialul global care comparat cu un threshold conduce la decizia de a considera tokenul sensibil sau nu.

In fine, sistemul are un Submodul de Postprocesare (SP2) care are rolul de a returna entitatile identificate, in formatul dorit si activarea functiei de lesire Date.

Transcrierea convorbirii in timp real poate fi fundamentala pentru operatiunea interna a aplicatiei de a anonimiza ireversibil conversatiile. Anonimizarea datelor ireversibil urmareste sa elimine necesitatea de a se obine consintamantul participantilor dintr-o conversatie, pentru ca aceasta sa fie inregistrata, conform legislatiei in viguare. Acest lucru este posibil, deoarece datele sunt sterse in timpul conversatiei si nu sunt salvate si procesate.

Este posibil ca datele anonimizate ireversibil obtinute de catre utilizatori sa fie inmagazinate in biblioteci mari de inregistrari pentru a-si imbunatati capacitatile de recunoastere si pentru a ajuta la dezvoltarea de noi produse care pot fi apoi oferite pietei in general.

Pentru a evita orice fel de reversibilitate a conversatiilor anonimizate nu se vor salva sau pastra documente istorice brute sau originale, datele originale fiind sterse automat in momentul in care au fost anonimizate in timpul conversatiei.

REVENDICARI

1. Sistem pentru anonimizarea datelor de identificare a persoanelor aflate intr-o convorbire audio/video , conectat la traficul dintre o Retea de Comunicatii (RC) si un Protocol de Comunicare (PC) printr-o Interfata de Programare a Aplicatiei (API) destinata interceptarii conversatiei careia i se aplica procedura de anonimizare a datelor si care ulterior anonimizarii, permite preluarea conversatiilor de catre un Subsistem de Transmitere (SDT) care le trimite dupa caz catre un Server de Analiza (SDA) si/sau un Server de Stocare (SDS), caracterizat prin aceea ca are in structura un Subsistem de Anonimizare al Conversatiilor(SAC) special conceput pentru inregistrarea, transcrierea discutiilor relevante provenite din arhitectura de comunicatii si anonimizarea ireversibil in timp real a datele cu caracter personal, fara ca datele utilizatorilor sa mai poata fi identificabile, sau implicit sa fie cerut acordul acestor utilizatori pentru inregistrarea discutiilor.
2. Sistem pentru anonimizarea datelor de identificare a persoanelor aflate intr-o convorbire audio/video conform revendicarii 1, caracterizat prin aceea ca Subsistemul de Anonimizare a Conversatiilor (SAC) este activat de functia Start automat in momentul in care Protocolului de Comunicare (PC) primeste un trafic de pachete si este alcatuit dintr-un Modul de Captare al Traficului (MCT) care intercepteaza si capteaza conversatia, automat, fara a fi necesara interventia altui modul, un Modul de Inregistrare Temporara al Conversatiei (MTTC) care are scopul de a pastra datele originale pana acestea sunt identificate si anonimizate, un Modul de Transcriere in Date(MTD), care poate transcrie conversatiile de exemplu in date audio reprezentate prin unde, valuri, date meta sau alte date caracteristice formatului audio, date text reprezentate de un transcript scris in cuvinte, fara a fi limitate de acestea, un Modul de Detectare al Datelor din Conversatie (MDDC) care urmareste sa gaseasca date cu caracter personal, dar nu limitate doar de acestea, care pot duce la identificarea participantilor din cadrul conversatiei si are rolul de a asigura ca toate datele acestora sunt anonimizate ireversibil, un Submodul de Anonimizare Ireversibila al Datelor (SAID), care folosind procedeuri diferite pentru tipuri diferite de date, astfel incat datele originale sa nu mai poata fi acesate si implicit sa nu poata duce la identificarea participantilor din discutie, un Modul de Stergere al Datelor Temporare (MSDT) activat dupa ce datele originale sunt anonimizate si informatiile relevante sunt sterse, un Modul de Transmitere(MT) cu rol de a transmite in timpul discutie sau la terminarea acesteia, toate datele fiind colectate dupa caz de catre un Modul de Analiza din Serverul de Analiza (MASDA) si sau un Modul de Stocare din Serverul de Stocare (MSSDS), si respectiv un Modul Incheiere Discutie

(MID) care detecteaza automat cand Protocol de Comunicare (PC) nu mai primeste trafic de pachete.

3. Metoda pentru anonimizarea datelor cu caracter personal a persoanelor aflate intr-o convorbire audio/video, implementata pe sistemul de la revendicarea 1 si 2, caracterizata prin aceea ce presupune parcurgerea urmatorilor pasi functionali :

- Pas.1 - se primeste prin protocolulul de comunicare (PC) traficul de pachete de date supus analizei ;
- Pas 2 - se intercepteaza si capteaza conversatia in mod automat ;
- Pas 3 - conversatia se transcrie dupa caz in date de tip audio, date de tip text sau similar ;
- Pas 4 - conversatia se inregistreaza temporar in scopul pastrarii datele originale pana acestea sunt identificate si anonimizate ;
- Pas 5 - se detecteaza datele din conversatie prin gasirea datelor cu caracter personal, dar nu limitate doar de acestea, care pot duce la identificarea participantilor din cadrul conversatiei, avand rolul de a asigura ca toate datele acestora vor fi anonimizate ireversibil ;
- Pas 6 - se realizeaza anonimizarea ireversibila a datelor utiliand procedeuri diferite pentru tipuri diferite de date, astfel incat datele originale sa nu mai poata fi acesate si implicit sa nu poata duce la identificarea participantilor din discutie ;
- Pas 7 - se sterg datele temporare si informatiile originale sunt sterse ;
- Pas 8 - toate datele colectate sunt transmise in timpul discutiei spre analiza si/sau dupa caz spre stocare ;
- Pas 9 - se detecteaza automat situatia in care sistemul nu mai primeste trafic de pachete, astfel rezultand ca discutia a fost incheiata.

4. Metoda pentru anonimizarea datelor cu caracter personal a persoanelor aflate intr-o convorbire audio/video conform revendicarii 3 caracterizata prin aceea ca in scopul implementarii pasului 6 destinat realizarii anonimizarii ireversibile a datelor, se parcurg urmatoarele etape procedurale :
-datele transcrise sunt preprocesate printr-un proces de segmentare (PS) cu rolul de segmentare a fiecarei propozitii pentru ca aceasta sa fie procesata individual, fara sa depinda de contextul general al conversatiei ;
-derularea unui proces de tokenizare (PT) care imparte datele prin parametrizare in unigrame, astfel incat tokenizarea cuvintelor individuale este reprezentata de spatiul regasit dupa fiecare cuvant transcris individual, iar seturile de cuvinte sunt reprezentate prin schimbarea vorbitorului ;

-derularea unui proces de analiza morfosintactica (PAM) care analizeaza unigramele, pentru a le clasifica si eticheta la nivel semantic pentru a oferi o structura a accestor etichete sub forma clasificata'

- derularea unui proces contectat paralel ponderat, in care ponderile pot sa fie setate in functie de cazul de utilizare. Fiecare din urmatoarele trei module calculeaza o probabilitate ca un token sa fie sau nu informatie personala care trebuie anonimizata :

- derularea unui proces compus dintr-un model bazat pe reguli (MBR) care implementeaza diferite reguli pentru a descoperi unele clase de entitati, care pot fi asociate cu date sensibile legate de categoria numere, aplicandu-se dupa caz o validare suplimentara efectuata pe toate numerele personale in care exista o validare de control, cifra de verificare sau suma de verificare, ceea ce permite dezambiguizarea si cresterea certitudinii cazurilor ;
- derularea unui proces de model bazat pe lexicon (MBL) care combina rezultatele procesului de analiza morfosintactica (PAM), un set de lexicoane si tehnici de derivatie si lematizare, cu scopul recunoasterii claselor de entitati, respectiv date cu caracter personal dar nu limitate de acestea ;
- derularea unui proces de model bazat pe invatare automata (MBIA) pentru validarea proceselor anterioare si descoperirea claselor de entitati cu o ambiguitate mai ridicata sau pentru cazurile in care nu exista reguli bine definite, cu rezultate maxime datorita retelelor neuronale ;

- derularea unui proces de functii cu rolul de a stabili suma ponderata a probabilitatilor care poate fi comparat cu un threshold.

-derularea unei postprocesari cu rolul de a returna entitatile identificate, in formatul dorit si activarea functiei de iesire date.

5. Produs program inregistrat pe un mediu citibil de catre un calculator, si care atunci cand este rulat pe calculator este destinat implementarii metodei pentru anonimizarea datelor cu caracter personal a persoanelor aflate intr-o convorbire audio/video, conform revendicarii 3 si 4.

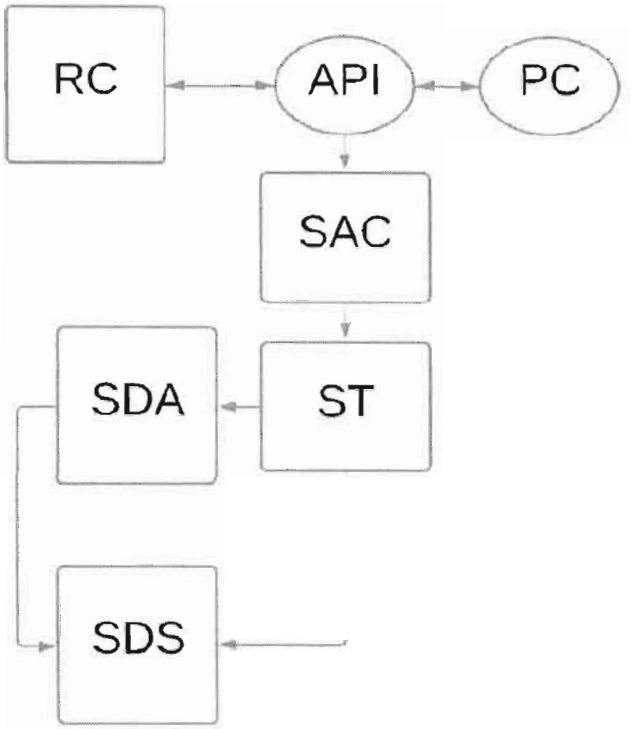

Fig.1

Fig.2

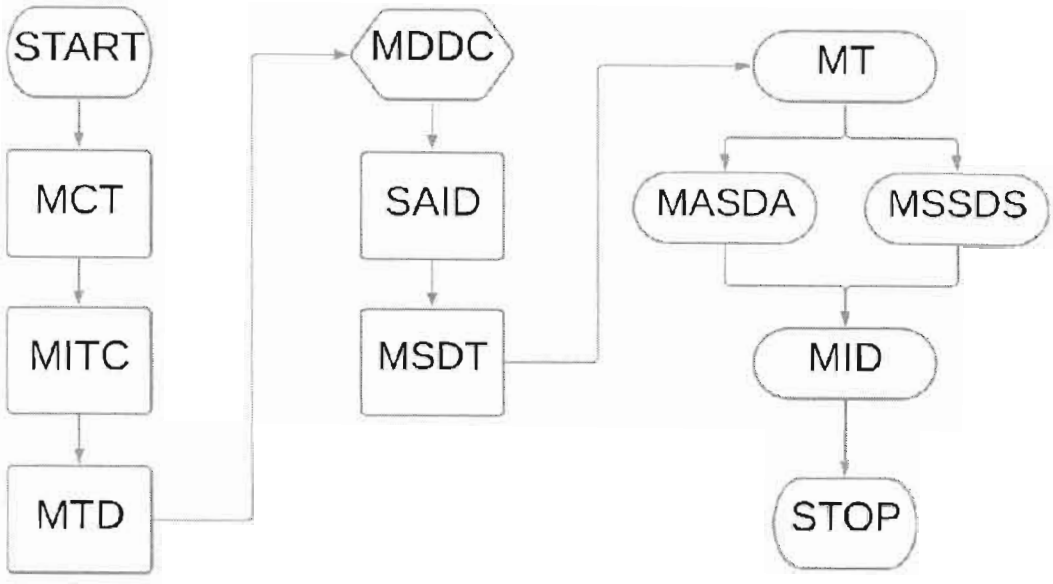
START
MCT
MITC
MTD
MDDC
SAID
MSDT
MT
MASDA
MSSDS
MID
STOP

Fig. 3

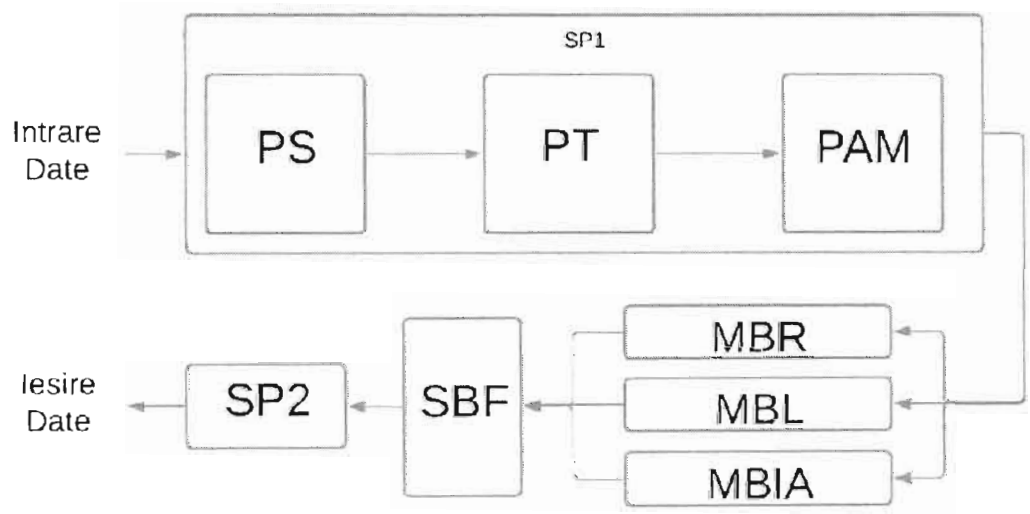
SP1
Intrare
Date
PS
PT
PAM
Iesire
Date
SP2
SBF
MBR
MBL
MBIA

Fig. 4