Fərdlərin Nitqlərinin Terminoloji Analizi Və İdentifikasiyası

Əfruz Qurbanova[1], Mehriban Bağırova[2]
[1,2]İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*afruz1961@gmail.com*, [2]*mehriban.amea@gmail.com*

Xülasə— **Məqalədə kiberkriminalistika kontekstində fərdlərin nitq xüsusiyyətlərinin identifikasiyası məsələləri araşdırılır, mövcud identifikasiya metodlarının terminoloji təhlili aparılır və fərdlərə xas olan fonetik, leksik və sintaktik xüsusiyyətlər də daxil olmaqla, nitq xüsusiyyətlərinin təsnifatı istiqamətində müasir yanaşmalar analiz olunur. Ənənəvi statistik modellərlə yanaşı, müasir süni intellekt və maşın təlimi texnologiyalarının tətbiqi imkanları göstərilir. Fərdlərin nitqlərinin terminoloji analizində mövcud ziddiyyətləri həll edəcək vahid terminoloji informasiya sisteminin yaradılmasının zəruriliyi əsaslandırılır. Məqalədə linqvistik analiz və hesablama metodlarının inteqrasiyasının rəqəmsal sübutların obyektiv qiymətləndirilməsində və kibercinayətkarlığın qarşısının alınmasında əhəmiyyəti verilir.**

Açar sözlər— ***kriminalistik linqvistika; terminoloji analiz; fərdin identifikasiyası; fonoskopiya; idiolekt; nitq xarakteristikası; süni intellekt.***

I. GİRİŞ

Müasir kriminalistikada cinayətkarın şəxsiyyətini müəyyən etmək üçün istifadə olunan ənənəvi metodlarla yanaşı, linqvistik analiz xüsusi çəkiyə malikdir. Hər bir fərdin nitqi onun "dil barmaq izi" (idiolekt) hesab olunur [1]. Xüsusilə anonim hədə-qorxu məktubları, kibercinayətkarlıq və terrorçuluqla mübarizədə nitqin terminoloji analizi həlledici rol oynayır. İnformasiya texnologiyalarının sürətli inkişafı kibercinayətkarlığın yeni formalarının yaranmasına və bu cinayətlərin araşdırılması ilə məşğul olan kiberkriminalistika sahəsinin inkişafına səbəb olmuşdur [2].

Kiberkriminalistikada fərdlərin nitqlərinin terminoloji analizi və identifikasiyası rəqəmsal mühitdə (sosial şəbəkələr, elektron poçt sistemləri, forum platformaları və digər onlayn ünsiyyət vasitələri) şəxslərin yazılı və ya şifahi ifadə tərzinin öyrənilməsi əsasında onların kimliyinin müəyyən edilməsi prosesidir. Bu istiqamət müasir kriminalistikanın sürətlə inkişaf edən sahələrindən biri olmaqla rəqəmsal sübutların linqvistik interpretasiyasına əsaslanır [3].

Kibercinayətkarlar çox vaxt anonimliyə güvənirlər, lakin onların "nitq imzası" rəqəmsal izdir, onu saxtalaşdırmaq və ya gizlətmək olduqca çətindir. Fərdin nitqində peşə terminlərindən, spesifik arqo və jarqonlardan istifadə etməsi onun sosial mənsubiyyətini və təhsil səviyyəsini daraldaraq identifikasiya olunması prosesini sürətləndirir. Sosial şəbəkələr, elektron yazışmalar və digər rəqəmsal mətnlərdə istifadə olunan leksik və sintaktik xüsusiyyətlər müəllifin sosial və peşəkar profilinin müəyyənləşdirilməsində əlavə informasiya mənbəyi rolunu oynayır.

Davam edən rəqəmsal transformasiya prosesləri cinayətkarlığın strukturunda keyfiyyət dəyişikliklərinə səbəb olmuşdur. Müasir cinayətlər virtual mühitdə törədilir, burada şifrələmə vasitələrinin, proksi serverlərin və anonimləşdirmə sistemlərinin istifadəsi ənənəvi identifikasiya metodlarını (məsələn, IP ünvanı və ya geolokasiya) təsirsiz hala gətirir [4]. Bu kontekstdə kompüter elmləri və məhkəmə linqvistikası metodlarını birləşdirən kiberkriminalistika ön plana çıxır. İnsanın nitq fəaliyyəti - istər mesajlaşmada yazılı mətn, istərsə də IP telefoniyada səs trafiki - fərdi tipoloji xüsusiyyətlərə malikdir. Bu fəaliyyətin terminoloji təhlili təhdid növünü təsnif etməklə yanaşı, həm də subyekti onun fərdi üslubu - idiolekt vasitəsilə müəyyən etməyə imkan verir [3]. Əsas məqsəd rəqəmsal mühitdə nitqin təhlilinə yanaşmaları sistemləşdirmək və mövcud identifikasiya metodlarının məhkəmə-ekspert praktikasında tətbiqi imkanlarını təhlil etməkdir. Tədqiqat işində fərdlərin nitqlərinin terminoloji analizi və identifikasiyası üçün statistik metodlar və süni intellekt modellərinin inteqrasiyası əsasında hibrid yanaşma təklif olunur.

II. KRİMİNALİSTİKA KONTEKSTİNDƏ FƏRDLƏRİN NİTQLƏRİNİN TERMİNOLOJİ ANALİZİ

Kriminalistik linqvistika (məhkəmə dilçiliyi) cinayətlərin açılmasında "dil barmaq izi"ndən istifadə edir. Hər bir fərdin nitqi onun sosial mənsubiyyəti, təhsili, peşəsi və psixoloji vəziyyəti haqqında unikal məlumatlar daşıyır. Kriminalistikada nitqin analizi müasir cinayət-prosessual araşdırmalarda ən mühüm sübut bazalarından biri hesab olunur [5]. Belə ki, naməlum şəxsin nitq nümunəsini şübhəli şəxslə müqayisə edərək eyniləşdirməklə, onu identifikasiya etmək olar. Eyni zamanda, nitqə əsasən müəllifin yaşını, cinsini, etnik mənşəyini və peşə sahəsini müəyyən etmək olar. İnformasiya texnologiyalarının inkişafı ilə səs yazıları, videomüşahidə və rəqəmsal mesajlar cinayətlərin açılmasında həlledici rol oynayır [6]. Cinayətkarın kimliyini müəyyən etmək üçün ən effektiv vasitələrdən biri onun həm şifahi, həm də yazılı ünsiyyətini analiz etməkdir. Aydındır ki, müasir informasiya texnologiyaları vasitəsilə şifahi nitq asanlıqla mətnə çevrilə bilir.

Terminoloji analiz: peşə və sosial "ləqəblər". Terminoloji analiz fərdin lüğət ehtiyatında üstünlük təşkil edən xüsusi sözlərin tədqiqidir. Bu, şəxsin peşə portretini yaratmağa kömək edir. Fərdin nitqinin terminoloji analizi nəticəsində aşağıdakılar aşkarlanır:

- *İxtisas terminləri:* məsələn, "IP ünvanı", "haker

DOI: 10.25045/forensic.gov.2026.37

hücumu" kimi terminlər şəxsin informasiya texnologiyaları sahəsinə, "qastrit", "anamnez" kimi sözlər isə tibb sahəsinə aid olduğunu göstərir.

- *Arqo və jarqonlar:* Kriminal aləmə məxsus jarqonların (məsələn, "şifrəli" danışıqlar) analizi cinayətkar qruplaşmaların daxili iyerarxiyasını üzə çıxarır.
- *Terminlərin yanlış istifadəsi:* Bəzən cinayətkarlar özlərini başqa peşə sahibi kimi qələmə verirlər. Lakin terminlərin qeyri-dəqiq və ya yerində işlədilməməsi onların "saxta profilini" ifşa edir.

Fərdin identifikasiyası zamanı nitqin aşağıdakı səviyyələrinə də baxılır:

- *Leksik-semantik səviyyə:* Tez-tez istifadə olunan "parazit" sözlər vasitəsilə fərd identifikasiya olunur. Regional dialektizmlər isə şəxsin haralı olduğunu müəyyən etməyə kömək edir.
- *Sintaktik səviyyə:* Cümlə quruluşuna görə, yəni cümlənin qısa və lakonik, yoxsa mürəkkəb və rəsmi-işgüzar olmasının müəyyən olunması, eyni zamanda, yazılı nitqdə durğu işarələrinin spesifik istifadəsi ilə fərd identifikasiya olunur.
- *Fonetik və akustik səviyyə (şifahi nitq üçün):* Səsin tembri, tezliyi və vurğu xüsusiyyətləri, eyni zamanda, nitqdəki fasilələr və tərəddüd anları fərdin identifikasiyasına kömək edir.

Səs yazılarının autentikliyinin yoxlanılması. Rəqəmsal manipulyasiyaların (montaj, "deepfake" audio) artdığı bir dövrdə nitqin analizinin texniki tərəfi ön plana çıxır. Belə ki, yazıya kənar müdaxilələrin olub-olmaması, səsyazma cihazının növünün və ətraf mühitin təyin edilməsi, fon küyünün analizi vasitəsilə hadisənin baş verdiyi məkanın (məsələn, küçə, qapalı otaq) identifikasiyası da əsas məsələlərdəndir.

III. METODOLOGİYA VƏ TEXNOLOGİYALAR

İdentifikasiya prosesi mətn məlumatlarının normallaşdırılması, linqvistik xüsusiyyətlərin çıxarılması və statistik modelləşdirmə mərhələlərindən ibarətdir. Müəllif davranışını xarakterizə edən leksik, sintaktik və morfoloji göstəricilər avtomatik analiz edilərək fərdin linqvistik profili formalaşdırılır. Əsas hesablama metodlarına N-gram statistikası, stilometrik analiz [7] və təbii dilin emalı (NLP, Natural language processing) əsaslı dərin təlim metodları daxildir. N-gram modelləri yazı ardıcıllıqlarını ölçərək müəllifə xas dil stereotiplərini müəyyən edir, stilometriya isə leksik zənginlik və struktur göstəriciləri əsasında müəllifliyin qiymətləndirilməsini həyata keçirir [8]. Müasir böyük dil modelləri (LLM, Large Language Models) mətnin semantik və kontekstual xüsusiyyətlərini analiz edərək identifikasiya dəqiqliyini artırır.

Hesablama metodlarının inteqrasiyası nəticəsində formalaşan linqvistik profil anonim müəlliflərin aşkarlanması və kibercinayətlərin analitik rekonstruksiyası üçün effektiv alət kimi çıxış edir. Müasir kriminalistikada nitqin analizi həm ekspert tərəfindən, həm də riyazi metodlar vasitəsilə aparılır. Bu yanaşmalar fərdin nitq nümunələrindən unikal linqvistik xüsusiyyətləri çıxarmağa və müəllifliyi dəqiqliklə müəyyən etməyə imkan verir. Nitqin identifikasiyası prosesi terminoloji analiz, statistik qiymətləndirmə və linqvistik modellərin inteqrasiyası ilə həyata keçirilir. Bu yanaşma hüquqi və texniki ekspertizalarda tətbiq olunur və rəqəmsal sübutların obyektiv qiymətləndirilməsini təmin edir.

IV. FƏRDLƏRİN NİTQLƏRİNİN İDENTİFİKASİYASI METODLARININ MÜQAYİSƏLİ ANALİZİ

Müasir kiberkriminalistikada linqvistik analiz metodlarının tətbiqi həm ənənəvi statistik modellər, həm də qabaqcıl süni intellekt alqoritmləri baxımından təhlil olunur:

Zipf Qanunu və leksik zənginlik. Nitqin identifikasiyasında ən məşhur statistik metodlardan biri Zipf qanunudur [9]. Zipf qanunu böyük verilənlər dəstində istənilən elementin tezliyinin həmin elementin tezlik cədvəlindəki sırası ilə tərs mütənasib olduğunu ifadə edən empirik bir prinsipdir. Bu qanun, ən çox rast gəlinən elementin ikinci ən çox rast gəlinən elementdən təxminən iki dəfə, üçüncü ən çox rast gəlinən elementdən üç dəfə və s. çox rast gəlindiyini bildirir. Hər bir fərdin özünəməxsus "söz tezliyi lüğəti" olur. Cinayətkarın yazdığı anonim məktubla şübhəli şəxsin əvvəlki yazıları arasındakı söz tezliyinin paylanması müqayisə edilir. Bu aşağıdakı kimi təyin olunur:

$$f(r) \approx \frac{C}{r} \quad (1)$$

Burada f(r) - sözün tezliyini, r- ranqını, C isə sabit əmsalı göstərir.

Kiberkriminalistika kontekstində Zipf modeli fərdi dil profilinin qurulması, müəllifliyin identifikasiyası zamanı leksik göstəricilərin seçilməsi üçün istifadə olunur.

Heaps qanunu. Heaps qanunu mətnin həcmi ilə lüğət tərkibinin artım dinamikası arasındakı əlaqəni ifadə edən empirik qanunauyğunluqdur. Fərdlərin leksik zənginliyinin ölçülməsi və müəllifliyin identifikasiyası üçün kiberkriminalistik analizdə geniş tətbiq olunur [6, 10]. Metodun riyazi ifadəsi aşağıdakı kimidir:

$$V(n) = K \cdot n^{\beta} \quad (2)$$

burada V- korpusda unikal və bir-birindən fərqlənən sözlərin sayı, n – sənəddə ümumi sözlərin (tokenlər) sayı, K və β - empirik təyin olunan sabit parametrlərdir, K adətən $10 < K < 100$ aralığında olur, β - adətən $0.4 < \beta < 0.6$ aralığında olur, leksik müxtəlifliyin artım sürətini göstərir. β-nın yüksək qiyməti daha zəngin lüğət ehtiyatını, aşağı qiyməti isə məhdud və təkrarlanan leksikanı xarakterizə edir.

Bayes teoremi. Kriminalistikada şübhəli şəxsin müəyyən bir mətni yazma ehtimalını hesablamaq üçün Bayes metodundan geniş istifadə olunur [11]. Metodun riyazi ifadəsi:

$$P(H \backslash E) = \frac{P(E \backslash H) \cdot P(H)}{P(E)} \quad (3)$$

Burada *P(H/E)* - verilmiş mətnin müəyyən müəllifə *(H)* aid olma ehtimalını, *P(E/H)* - həmin müəllifə xas dil xüsusiyyətlərinin müşahidə olunma ehtimalını, *P(H)* - ilkin (aprior) ehtimalı, *P(E)* isə müşahidə olunan sübutların ümumi ehtimalını ifadə edir.

N-qram analizi. N-gram metodu mətndə ardıcıl gələn söz

qruplarının tezliyini ölçür. Bu, fərdin sintaktik vərdişlərini (şüuraltı yazı tərzini) üzə çıxarır [12]. Bi-gram və tri-gram modelləri vasitəsilə müəllifin spesifik leksik və sintaktik vərdişləri modelləşdirilir. Əgər anonim hədə-qorxu məktubunda "vaxt, çatdı" və ya "cavab, gözləyirəm" bi-gramları statistik olaraq normadan çox təkrarlanırsa, bu, şübhəli şəxslərin yazı bazası ilə müqayisə edilir. Riyazi olaraq Kullback-Leibler (iki ehtimal paylanması arasındakı fərq) məsafəsi ilə bu iki yazı tərzinin eyni şəxsə aid olub-olmaması müəyyən edilir [13].

Klaster analizi. Mətnlərin eyni şəxs tərəfindən yazılıb-yazılmadığını müəyyən etmək üçün iyerarxik klasterləşdirmə metodu tətbiq edilir. Bu yanaşma kiberkriminalistikada anonim müəlliflərin ilkin seqmentasiyası, müəllif profilinin müəyyənləşdirilməsi və verilənlər bazasındakı anomaliyaların aşkarlanması üçün effektiv vasitə hesab olunur [14]. Mətnlərdəki terminlərin sıxlığı, cümlə uzunluqları və durğu işarələrinin istifadəsi kimi parametrlər koordinat sistemində nöqtələr şəklində qeyd olunur. Bir-birinə yaxın olan (riyazi məsafə baxımından – Evklid məsafəsi) nöqtələr bir "klasterdə" birləşir. Əgər şübhəli şəxsin yazısı naməlum mətnlə eyni klasterə düşürsə, bu, identifikasiya üçün güclü sübutdur.

Diskriminant analizi. Diskriminant analizi fərdləri müəyyən qruplara (məsələn: yaş qrupları, təhsil səviyyəsi və ya peşə sahələri) ayırmaq üçün istifadə olunan çoxölçülü statistik klassifikasiya metodudur və mətnlərin linqvistik və stilometrik xüsusiyyətləri əsasında müəllifin identifikasiyasını həyata keçirməyə imkan verir. Bu yanaşma anonim mətnlərin təhlili, bot və süni intellekt tərəfindən yaradılmış mətnlərin aşkarlanması, həmçinin kiberkriminalistikada rəqəmsal sübutların müqayisəli analizində geniş tətbiq olunur [15]. Fərdin rəqəmsal nitqinin leksik zənginliyini və terminoloji sıxlığını müəyyən etmək üçün Type-Token Ratio (TTR) dəyəri istifadə olunur. Type-Token Ratio leksik müxtəlifliyin sadə ölçüsüdür, unikal sözlərin (növlərin) sayının ümumi sözlərin (işarələrin) sayına nisbətini hesablayır [16]:

$$TTR = \text{Number of Types (unique words)} / \text{Number of Tokens (total words)}$$

Əmsallar [0,1] aralığında dəyişir. Bu göstərici linqvistikada və dil qiymətləndirməsində geniş istifadə olunur, lakin mətnin uzunluğundan çox asılıdır, çünki daha uzun mətnlərdə TTR göstəricisi daha aşağı olur.

Süni intellekt və maşın təlimi metodları. Müasir kiberkriminalistikada fərdin nitqini identifikasiya etmək üçün artıq sadəcə statistik cədvəllər kifayət etmir, alqoritmlər nitqin çoxsaylı xüsusiyyətlərini eyni vaxtda təhlil edir. Ənənəvi N-gram və ya LSTM (Long Short-Term Memory) modellərindən Transformer arxitekturasına keçid, mətndəki "müəllif barmaq izi"nin aşkarlanmasında keyfiyyətcə yeni mərhələdir. Transformer modelləri (BERT, RoBERTa, GPT seriyası) "özünə-diqqət" (Self-Attention) mexanizmindən istifadə edərək, cümlədəki hər bir sözün digər sözlərlə olan əlaqəsini məsafədən asılı olmayaraq eyni anda qiymətləndirir.

Vektorlaşdırma (Word Embeddings). Süni intellekt hər bir sözü çoxölçülü fəzada rəqəmsal vektora çevirir. Word2Vec və ya BERT modelləri sözlərin təkcə özünü deyil, onların əhatə olunduğu konteksti də analiz edir [17]. Cinayətkarın terminləri hansı kontekstdə işlətməsi (məsələn, "paket" sözünü texnoloji, yoxsa narkotik kontekstində işlətməsi) onun peşə və sosial mühitini 99% dəqiqliklə müəyyən etməyə imkan verir. Klassik "Word2Vec" metodlarından fərqli olaraq, Transformer modelləri sözlərin çoxmənalı (polisemantik) təbiətini dərk edir. Məsələn, cinayətkarın "iş" sözünü həm texniki, həm də kriminal kontekstdə istifadə etməsi Transformer tərəfindən çoxölçülü vektor fəzasında fərqli koordinatlarla qeyd edilir. Bu, müəllifin lüğət ehtiyatının və semantik yanaşmasının dəqiq profilini çıxarmağa imkan verir.

Neyron şəbəkələri (LSTM və Transformers). LSTM (Long Short-Term Memory) şəbəkələri mətndəki uzunmüddətli asılılıqları yadda saxlayır. İnsan yazarkən cümlənin əvvəli ilə sonu arasında özünəməxsus bir ritm və məntiq qurur. Alqoritm şübhəli şəxsin yazılarındakı bu "ritmik qəlibi" öyrənir və anonim mətndə həmin qəlibin izini axtarır. LSTM modelləri mətndəki uzun məsafəli əlaqələri saxlamaqda çətinlik çəkirdi. Transformer modelləri isə bütün mətni "görə bildiyi" üçün, bir səhifəlik mətnin əvvəlindəki terminlə sonundakı sintaktik struktur arasındakı gizli əlaqəni tuta bilir. Bu, böyük həcmli rəqəmsal sübutların təhlilində identifikasiya dəqiqliyini əhəmiyyətli dərəcədə artırır.

Dəstək Vektor Maşınları (SVM - Support Vector Machines). SVM təsnifat alqoritmidir. Əgər əlimizdə iki fərqli şübhəli şəxs varsa, SVM onların nitq xüsusiyyətləri arasında ən optimal ayırıcı xətti (hyperplane) çəkir [18].

Süni intellekt alqoritmləri böyük həcmli mətn korpuslarını qısa müddətdə emal edə bilir. İnsan faktorundan asılı olan subyektiv rəylər (məsələn, "məncə bu yazı tərzi həkimə bənzəyir") riyazi dəqiqliklə əvəz olunur. Ənənəvi analiz üsulları ilə aşkarlanması çətin olan dil xüsusiyyətləri (durğu işarələrindən sonrakı boşluqların sayı, bağlayıcıların istifadə tezliyi və s.) alqoritmlər tərəfindən aşkar edilir.

Qiymətləndirmə və etibarlılıq. Kiberkriminalistikada süni intellekt modellərinin məhsuldarlığı "dəqiqlik" (accuracy) faizi ilə yanaşı tamlıq (Recall), F1-score ilə də ölçülməlidir. Bu səbəbdən, modelin etibarlılığı aşağıdakı metrikalarla qiymətləndirilməlidir:

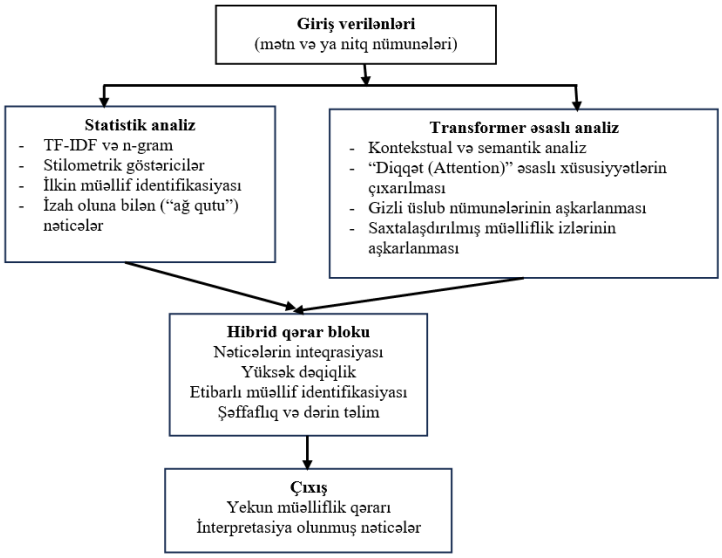
- *Dəqiqlik (Precision):* Modelin "cinayətkar" olaraq işarələdiyi şəxslərin neçə faizi həqiqətən cinayətkardır?
- *Tamlıq (Recall):* Həqiqi cinayətkarların neçə faizi model tərəfindən aşkar edilib?
- *F1-Score:* Dəqiqlik və tamlıq arasındakı harmonik ortadır. Kriminalistika kimi balans tələb edən sahələrdə modelin ən yaxşı "tarazlıq" göstəricisidir.

Modelin nəticələrini vizuallaşdırmaq üçün qarışıqlıq matrisi (Confusion Matrix) tətbiq olunur. Bu matris hansı növ səhvlərə yol verildiyini göstərir. ROC-AUC (Receiver Operating Characteristic-Area Under Curve) əyrisi isə modelin müxtəlif həssaslıq səviyyələrində nə dərəcədə stabillik göstərdiyini müəyyən edir. AUC dəyərinin 1-ə yaxın olması, modelin təsadüfi proqnozdan kəskin fərqlənən yüksək etibarlıq qabiliyyətinə malik olduğunu sübut edir. Cədvəl I-də kiberkriminalistika kontekstində fərdlərin identifikasiyası metodlarının müxtəlif meyarlar üzrə müqayisəli analizi verilmişdir.

CƏDVƏLI. KRİMİNALİSTİKA KONTEKSTİNDƏ MÜƏLLİFLİYİN İDENTİFİKASİYASI METODLARININ MÜQAYİSƏLİ TƏHLİLİ

Meyarlar	**Ənənəvi statistik metodlar (Zipf, Heaps, TTR)**	**Transformer modelləri (BERT, GPT, RoBERTa)**
Təhlil səviyyəsi	Leksik (söz tezliyi, lüğət zənginliyi)	Semantik və sintaktik (kontekst, struktur)
Kontekstual dərk etmə	Məhdud (sözlərin ardıcıllığına baxılır)	Yüksək (cümlə daxili və cümlələrarası əlaqə)
İzah edilə bilmə	Yüksək (şəffaf riyazi düsturlar)	Aşağı (qara qutu, XAI tələb edir)
Resurs tələbatı	Çox aşağı (adi hesablama gücü)	Yüksək (GPU/TPU və böyük data tələb edir)
Məhkəmə sübutlarının etibarlılığı	Məntiqi əsaslandırması asandır	İzah edilə bilən (XAI) vasitəsilə təsdiqlənir
İnsan faktorundan asılılıq	Aşağı	Çox aşağı (tam avtomatlaşdırılmış)
İlkin verilənlər tələbi	Az həcmli mətnlər üçün uyğundur	Böyük həcmli təlim məlumatları tələb edir

Bu müqayisədən göründüyü kimi, hər iki yanaşmanın özünəməxsus üstünlükləri var. Statistik metodlar məlumatın az olduğu və ya dərhal nəticə tələb olunan ilkin tədqiqat mərhələlərində əhəmiyyətlidir. Onların "ağ qutu" (white box) xarakteri məhkəmədə ekspertin izahat verməsini asanlaşdırır. Transformer modelləri isə mürəkkəb kibercinayətlərdə, mətnin saxtalaşdırıldığı və ya cinayətkarın üslubunu gizlətməyə çalışdığı hallarda "Diqqət (Attention)" mexanizmi sayəsində gizli nümunələri aşkar etmək üçün daha effektlidir. Beləliklə, müqayisəli analiz göstərdi ki, kiberkriminalistika kontekstində fərdlərin nitqlərinin terminoloji analizi və identifikasiyasında "Hibrid yanaşma" (Hybrid approach) məqsədəuyğundur. Belə ki, ilkin identifikasiya üçün statistik metodlardan, dərin analiz və təsdiqləmə üçün isə Transformer modellərindən istifadə edən çoxpilləli bir sistem ən yüksək dəqiqliyi təmin edir (Şəkil 1.)



Şəkil 1. Müəllifliyin identifikasiyası üçün hibrid intellektual sistemin arxitekturu

İlkin mərhələdə xam verilənlər (məsələn, məqalələr, elektron məktublar və ya transkripsiya olunmuş nitq) sistemə daxil edilir. Sistem eyni vaxtda iki fərqli metodologiya ilə işləyir. TF-IDF (Term Frequency-Inverse Document Frequency) və n-gram modelləri vasitəsilə sözlərin işlənmə tezliyi ölçülür. Belə ki, stilometrik göstəricilər (cümlə uzunluğu, durğu işarələrinin istifadəsi və s.) vasitəsilə müəllifin yazı üslubu kəmiyyətlərlə ifadə olunur. Burada qərarın hansı statistik göstəriciyə əsasən verildiyini asanlıqla izah etmək mümkündür. Transformer modelləri əsasında "diqqət" mexanizmi vasitəsilə mətndəki sözlər arasındakı uzaq kontekstual əlaqələr müəyyən edilir. Burada müəllifin gizli, bəzən şüuraltı səviyyədə üslub nümunələrini və saxtalaşdırılmış müəlliflik izlərini (məsələn, başqasının üslubunu təqlid etmə cəhdlərini) aşkarlamaq üçün semantik analiz aparılır. Sonrakı addımda – qərarların birləşməsi prosesi həyata keçirilir. Statistik analizdən gələn dəqiq məlumatlarla, Transformer modelindən gələn dərin semantik nəticələr inteqrasiya olunur, şəffaflıq və dərinlik balanslaşdırılır, bu da sistemin həm yüksək dəqiqliklə işləməsini, həm də əsaslandırılmış qərarlar verməsini təmin edir. Sistemin çıxışında mətnin konkret olaraq kimə aid olduğu müəyyən edilir və hesabat alınır.

NƏTİCƏ

Aparılan tədqiqat göstərir ki, rəqəmsal transformasiya dövründə kiberkriminalistika sahəsində nitqin analizi yalnız köməkçi vasitə deyil, cinayətlərin açılmasında əsas komponentə çevrilmişdir. Tədqiqat nəticəsində məlum oldu ki, riyazi-statistik modellər (Zipf, Heaps, diskriminant analizi) fərdi dil profilinin təməlini təşkil edir, süni intellekt texnologiyaları (Neyron şəbəkələri, Transformer modelləri) isə bu analizi daha sürətli və dəqiq icra etməyə imkan verir. Ekspertlərin subyektiv qiymətləndirməsi yerini sübuta əsaslanan riyazi modellərə vermişdir. Bu, rəqəmsal sübutların məhkəmə-ekspert praktikasında etibarlılığını artırır. "İdiolekt"in (dil barmaq

izinin) rəqəmsal izləri, anonimləşdirmə sistemləri (VPN, proksi və s.) tətbiq edilsə belə, cinayətkarın sosial, peşəkar və psixoloji profilini müəyyən etməyə imkan verir. Fərdlərin nitqlərinin terminoloji analizində mövcud ziddiyyətləri həll etmək üçün vahid terminoloji informasiya sisteminin yaradılması və Milli terminoloji informasiya sisteminə [19] inteqrasiyası zəruridir.

Təklif olunan hibrid model əsasında statistik metodların stabilliyi ilə süni intellektin güclü analiz qabiliyyətinin birləşdirilməsi müəllifliyin identifikasiyasında maksimum effektivliyin əldə edilməsinə kömək edəcəkdir.

Tədqiqat göstərir ki, gələcək hüquqi ekspertiza fəaliyyəti linqvistik biliklərin kompüter elmləri ilə sıx inteqrasiyasından birbaşa asılıdır. Təsvir olunan yanaşmaların hüquq-mühafizə orqanlarının təcrübəsinə tətbiqi, kibercinayətkarlıqla mübarizədə daha çevik və effektiv nəticələrin əldə olunmasına şərait yaradacaqdır.

ƏDƏBİYYAT

[1] T. A. Litvinova, "Пунктуационные выборы как составляющая ортологического параметра идиолекта носителя современного русского языка в аспекте идентификационной автороведческой экспертизы," Политическая лингвистика, no. 1, pp. 114-121, 2019.

[2] A. V. Serebrennikova, "Криминологические проблемы цифрового мира (Цифровая криминология)," Всероссийский криминологический журнал, vol. 14, no. 3, pp. 423-430, 2020.

[3] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing ‐ style features and classification techniques," Journal of the American Society for Information Science and Technology, vol. 57, no. 3, pp. 378-393, 2006.

[4] C. Chen and B. Dong, "Digital forensics analysis based on cybercrime and the study of the rule of law in space governance," Open Computer Science, vol. 13, no. 1, pp. 1-12, 2023.

[5] M. Coulthard, "Author identification, idiolect, and linguistic uniqueness," Applied Linguistics, vol. 25, no. 4, pp. 431-447, 2004.

[6] E. Stamatatos, "A survey of modern authorship attribution methods," Journal of the American Society for Information Science and Technology, vol. 60, no. 3, pp. 538-556, 2009.

[7] M. Kestemont, "Function words in authorship attribution. From black magic to theory?," in Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL), 2014, pp. 59-66.

[8] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," Journal of the American Society for Information Science and Technology, vol. 60, no. 1, pp. 9-26, 2009.

[9] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," Psychonomic Bulletin & Review, vol. 21, no. 5, pp. 1112-1130, 2014.

[10] A. Hernández-Fernández, I. G. Torre, J. M. Garrido, and R. Ferrer-i-Cancho, "Linguistic laws in speech: the case of Catalan and Spanish," Entropy, vol. 21, no. 12, p. 173-188, 2019.

[11] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Machine Learning, vol. 29, no. 2, pp. 131-163, 1997.

[12] J. Houvardas and E. Stamatatos, "N-gram feature selection for authorship identification," in International Conference on Artificial Intelligence: Methodology, Systems, and Applications, Berlin, Heidelberg: Springer, 2006, pp. 77-86.

[13] F. Raiber and O. Kurland, "Kullback-Leibler Divergence Revisited," in Proceedings of the International Conference on Theory of Information Retrieval (ICTIR'17), Amsterdam, The Netherlands, 2017, pp. 117-124.

[14] M. Eder, J. Rybicki, and M. Kestemont, Stylometry with R: A Package for Computational Text Analysis, 2016, pp. 107-121.

[15] P. Kayarkar and P. Ricchariaya, "An enhanced approach for digital forensics using innovative gsp algorithm," International Journal of Computer Applications, vol. 103, no. 6, pp. 18-22, 2014.

[16] P. Rosillo-Rodes, M. San Miguel, and D. Sánchez, "Entropy and type-token ratio in giga word corpora," Physical Review Research, vol. 7, no. 3, pp. 1-14, 2025. https://journals.aps.org/prresearch/pdf/10.1103/rxxz-lk3n.

[17] C. D. Putra and H. C. Wang, "Advanced ERT-CNN for hate speech detection," Procedia Computer Science, vol. 234, pp. 239-246, 2024.

[18] B. Fakiha, "Enhancing Cyber Forensics with AI and Machine Learning: A Study on Automated Threat Analysis and Classification," International Journal of Safety & Security Engineering, vol. 13, no. 4, pp. 701-707, 2023.

[19] R. M. Alguliyev and A. M. Gurbanova, "The Conceptual Foundations of National Terminological Information System," International Journal of Education and Management Engineering (IJEME), vol. 8, no. 4, pp. 19-30, 2018. https://www.mecs-press.org/ijeme/ijeme-v8-n4/IJEME-V8-N4-3.pdf.

Terminological Analysis And Identification of the Speech of Individual Persons

Afruz Gurbanova[1], Mehriban Baghirova[2]

[1,2]Institute of Information Technology, Baku, Azerbaijan

[1]afruz1961@gmail.com, [2]mehriban.amea@gmail.com

Abstract- This article examines the identification of speech characteristics of individuals in the context of cybercrime. It provides a terminological analysis of existing identification methods and analyzes modern approaches to classifying speech characteristics, including phonetic, lexical, and syntactic features specific to individuals. Along with traditional statistical models, the potential for applying modern artificial intelligence and machine learning technologies is demonstrated. The need for a unified terminology information system is substantiated, which will resolve existing contradictions in the terminological analysis of individual speech. The article emphasizes the importance of integrating linguistic analysis and computational methods in the objective evaluation of digital evidence and the prevention of cybercrime.

Keywords- forensic linguistics; terminological analysis; individual identification; phonoscopy; idiolect; speech characteristics; artificial intelligence.