

Fərdi Dil Korpusları Əsasında Şəxslərin Çoxməqsədli İdentifikasiyası Problemləri və Həlli Yolları

Əfruz Qurbanova¹, Səriyə Qasimova²

¹Informasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan

²Naxçıvan Dövlət Universiteti, Naxçıvan, Azərbaycan

¹afruz1961@gmail.com, ²qasimovaseriyye@ndu.edu.az

Xülasə— Müasir rəqəmsal kommunikasiya mühitində fərdlərin yazılı nitq nümunələri şəxsi dil korpuslarının yaradılması üçün mühüm mənbəyə çevrilir. Məqalədə fərdi dil korpusları əsasında müəllifin identifikasiyası və profiləşdirilməsi üçün informativ əlamətlərin çıxarılması problemi araşdırılır. Tədqiqat işində leksik, morfoloji, sintaktik, diskurs-stilometrik, emosional, dialektal, frazeoloji-mədəni və oxunabilirlik göstəricilərini vahid çoxsəviyyəli profil vektorunda birləşdirən model təklif edilmişdir. Hər bir əlamətin informativliyi onun statistik əlaqə, sabitlik, interpretasiya dəyəri və profil kateqoriyası ilə uyğunluğu əsasında qiymətləndirilir. Təklif olunan model korpusun formalaşdırılması, ilkin emal, linqvistik annotasiya, əlamət bloklarının çıxarılması, informativlik indeksinin hesablanması, əlamətlərin seçilməsi və identifikasiya mərhələlərini əhatə edir. Bu yanaşma Azərbaycan dili üçün fərdi dil korpuslarının intellektual analizi, regional dil xüsusiyyətlərinin aşkarlanması və izah edilə bilən süni intellekt sistemlərinin hazırlanması baxımından perspektivlidir.

Açar sözlər— fərdi dil korpusu, müəllif profili, informativ əlamətlər, dialektometriya, emosional analiz, frazeoloji markerlər, oxunabilirlik

I. GİRİŞ

Rəqəmsal transformasiya şəraitində insanın dil fəaliyyəti getdikcə daha çox mətn şəklində qeydə alınır. Sosial şəbəkə paylaşımları, forum yazıları, bloqlar, elektron məktublar, şərhlər və müxtəlif rəqəmsal platformalarda yaranan mətnlər konkret şəxsin dil davranışının təhlilinə imkan verən dəyərli resurslardır. Bu mətnlərin müəyyən prinsiplərlə toplanaraq strukturlaşdırılması fərdi dil korpuslarının yaradılmasına imkan verir. Fərdi dil korpusu müəyyən bir müəllifə məxsus mətnlərin elektron toplusu olmaqla həmin şəxsin üslubi, sintaktik, semantik, regional, emosional və mədəni xüsusiyyətlərini araşdırmaq üçün mühüm elmi baza yaradır.

Müəllif profiləməsi sahəsində aparılmış tədqiqatlar göstərir ki, yazılı mətnlər müəllifin müxtəlif xüsusiyyətlərinin müəyyənəşdirilməsi üçün əhəmiyyətli mənbədir. Bu sahədə məzmun əsaslı və üslub əsaslı əlamətlərin birgə istifadəsi müəllifin kimliyinin və müəyyən xüsusiyyətlərinin təyində əsas yanaşmalardan biri hesab olunur [1]. Sonrakı işlər müəllif profiləməsinə linqvistik profiləmə istiqamətində genişləndirərək mətnlərdən müxtəlif səviyyəli dil göstəricilərinin çıxarılmasını mümkün edən alətlər və yanaşmalar təklif etmişdir [2, 3].

Fərdi dil korpuslarının araşdırılması yalnız müəllifin ümumi

profilinin qurulması ilə məhdudlaşmır. Belə korpuslar əsasında müəllifin peşə yönümü, emosional vəziyyəti, dil bilgisi səviyyəsi, ləhcə-bölgə mənsubiyyəti, folklor və frazeoloji fonda yaxınlığı, yazı keyfiyyəti və mətnin mürəkkəbliyi səviyyəsi barədə nəticələr əldə etmək mümkündür. Tədqiqatın məqsədi həmin göstəriciləri sistemləşdirən və identifikasiya üçün alqoritmik ardıcılıq verən çoxsəviyyəli informativ əlamətlərin çıxarılması metodunu əsaslandırmaqdır.

II. MÜƏLLİF PROFİLLƏMƏSİ VƏ LİNGVİSTİK PROFİLLƏMƏ

Müəllif profiləməsi yazılı mətnlər əsasında müəllifin müxtəlif xüsusiyyətlərinin avtomatik və ya yarımavtomatik təyin olunması problemini ifadə edir. Bu sahənin əsas ideyası ondan ibarətdir ki, mətn yalnız informasiya daşıyıcısı deyil, həm də onu yaradan şəxsin sosial, psixoloji və linqvistik cəhətlərini daşıyan sistemdir. Klassik yanaşmalarda müəllifə məxsus məndəki xüsusiyyətlər iki əsas qrupa bölünür: məzmun əsaslı əlamətlər və üslub əsaslı əlamətlər [1].

Bu sahədə aparılan tədqiqatlar göstərmişdir ki, müəllif profiləməsi yalnız söz tezliklərinə və ya məzmun oxşarlığına söykənməməlidir. Leksik, morfoloji, sintaktik, diskurs və stilometrik xüsusiyyətlərin birlikdə təhlili daha səmərəli nəticələr verir. Bu baxımdan linqvistik profiləmə müəllif profiləməsinin daha geniş və dərin forması kimi çıxış edir. Profiling-UD kimi alətlər mətnlərdən 130-dan artıq linqvistik göstərici çıxarmaqla müəllifin dilini müxtəlif səviyyələrdə karakterizə etməyə imkan verir [2-5].

Müasir müəllif profiləşdirilməsi tədqiqatları göstərir ki, gender, yaş, dil müxtəlifliyi, region, sosial fon və bəzi hallarda psixoloji xüsusiyyətlər mətnlər əsasında müəyyən dərəcədə proqnozlaşdırıla bilər [4, 6]. Bununla yanaşı, transformer əsaslı modellərin inkişafına baxmayaraq, interpretasiya qabiliyyəti yüksək olan linqvistik əlamətlərdən istifadə edən klassik və hibrid yanaşmalar hələ də mühüm əhəmiyyət daşıyır [7, 8].

III. FƏRDİ DİL KORPUSLARINDAN ÇIXARILAN ƏLAMƏT QRUPLARI

Fərdi dil korpuslarından çıxarılan əlamətlər çoxqatlı xarakter daşıyır. Leksik əlamətlər müəllifin söz seçimi, söz tezliyi, terminoloji sıxlığı və n-qram modelləri ilə bağlıdır. Müxtəlif tədqiqatlar göstərir ki, söz və simvol n-qramları müəllif profiləməsində ən çox istifadə olunan xüsusiyyətlərdəndir [1, 4, 6]. Bu göstəricilər müəllifin peşəkar yönümünü, mövzu sahəsini və qismən mədəni dil mühitini əks etdirə bilər.

Morfoloji və sintaktik xüsusiyyətlər daha dərin və nisbətən

sabit linqvistik təbəqəni əks etdirir. Buraya nitq hissələri ardıcılığı, şəkilçi və qrammatik formalar, sintaktik asılılıq əlaqələri, cümlələrin orta uzunluğu, mürəkkəb cümlələrin payı və sintaktik müxtəliflik daxildir. Sintaktik və diskurs əlamətlərinin müəllif identifikasiyasında əhəmiyyətini araşdıran tədqiqatlar göstərir ki, az sayda, lakin yaxşı seçilmiş sintaktik və diskurs xüsusiyyətləri yüksək nəticə verə bilər [5].

Diskurs və stilometrik xüsusiyyətlər müəllifin mətn təşkil etmə üsulunu, ardıcılıq və əlaqəlilik yaratma strategiyasını, diskurs markerlərindən istifadə verdiyini, durğu işarələri və funksional sözlərlə bağlı davranışını əks etdirir. Bu əlamətlər müəllifin fikri necə bağlamasını, səbəb-nəticə və qarşılaşdırma münasibətlərini necə qurmasını, arqumentasiyanı necə təşkil etməsini göstərən mühüm siqnallardır.

Yazı keyfiyyəti və oxunabilirlik göstəriciləri müəllifin dil səriştəsi barədə məlumat verir. Bu məqsədlə leksik zənginlik, söz növü müxtəlifliyi əmsali (Type-Token Ratio, TTR), sintaktik müxtəliflik, cümlə mürəkkəbliyi, qrammatik düzgünlük və oxunaqlıq ölçülərindən istifadə olunur [8]. Bu əlamətlər "intellektual səviyyə" anlayışını birbaşa ölçmək iddiası yaratmadan "koqnitiv-linqvistik mürəkkəblilik" və "mətnin struktur keyfiyyəti" kimi daha elmi parametrlərlə ifadə etməyə imkan verir.

Emosional əlamətlərə müsbət-mənfi ton, konkret emosiyaları bildirən leksik vahidlər, vurğulayıcı strukturlar, nida və təkrar konstruksiyalar daxildir. Emosiyaların aşkarlanması və sentiment analizi üzrə tədqiqatlar göstərir ki, emosional məzmunun təhlili üçün həm leksik, həm də kontekstual və struktur xüsusiyyətlərdən istifadə etmək mümkündür [9, 10].

Dialekt və bölgə mənsubiyyətini əks etdirən əlamətlər regional variasiya baxımından xüsusi əhəmiyyət daşıyır. Korpus əsaslı dialektometriya tədqiqatları regional fərqlərin leksik, morfosintaktik və tezlik əsaslı göstəricilər vasitəsilə ölçülə biləcəyini göstərmişdir [11]. Dialekt identifikasiyası üzrə tətbiqi işlərdə n-qrammlar, TF-IDF, vektor təqdimatı (embedding) və dərin təlim yanaşmalarından geniş istifadə edilir [12].

Frazeoloji vahidlər, atalar sözləri, deyimlər və digər sabit ifadələr mədəni identiklik və folklor fondunu əks etdirən mühüm göstəricilərdir. Frazeologiya və mədəni kimlik arasındakı əlaqəni araşdıran işlər bu vahidlərin kollektiv dəyərlərin, dünya görüşünün və mədəni yaddaşın daşıyıcısı olduğunu göstərir [13, 14].

IV. DIALEKTOMETRİYA, EMOSİONAL ANALİZ VƏ MƏDƏNİ MARKERLƏR

Dialektometriya dil variasiyasının kəmiyyətə ölçülməsi və coğrafi baxımdan təhlili ilə məşğul olan mühüm istiqamətdir. Korpus əsaslı dialektometriya yanaşması atlas məlumatlarından başqa təbii və autentik mətn korpuslarına söykənməyi də mümkün edir. Bu yanaşmada dialektlərarası məsafələrin hesablanması, məsafə matrisinin qurulması və regional münasibətlərin xəritələndirilməsi əsas mərhələlər kimi çıxış edir [11].

Regional müəllif profiləməsi üzrə son tədqiqatlar dialektometriyanı sosial media verilənləri ilə birləşdirir. Geolokasiyalı sosial media məlumatları əsasında regional

markerlərin çıxarılması və məkan statistikasını ilə onların yayılmasının təhlili müasir regional müəllif profilinin formalaşdırılmasına imkan verir [15]. Bu yanaşma ənənəvi dialektoloji metodları tam əvəz etmir, lakin onları kəmiyyət göstəriciləri və real istifadə kontekstləri ilə zənginləşdirir.

Emosional vəziyyətin mətnlərdən aşkarlanması NLP sahəsinin mühüm istiqamətlərindən biridir. Sentiment analizi əsasən müsbət, mənfi və neytral münasibəti müəyyən edirsə, emosiya aşkarlama qəzəb, sevinc, qorxu, kədər və digər konkret emosional halları müəyyən etməyə yönəlir [10]. Fərdi korpusların uzunmüddətli təhlili müəllifin davamlı emosional tonunu, qısamüddətli təhlili isə situativ emosional vəziyyətini xarakterizə edə bilər.

Folklor mənsubiyyətinin müəyyənəşdirilməsi üçün frazeoloji və paremioloji vahidlərin təhlili məhsuldar yanaşmadır. Atalar sözləri, deyimlər və frazeoloji vahidlər müəyyən dil icmasının kollektiv təcrübəsini, dəyərlər sistemini və mədəni yaddaşını əks etdirən sabit dil vahidləridir [13]. Bu baxımdan fərdi korpusda belə vahidlərin mövcudluğu, tipi, tezliyi və kontekstdəki funksiyası müəllifin mədəni-linqvistik profilinin formalaşdırılmasında informativ rol oynaya bilər.

V. ÇOXSƏVİYYƏLİ INFORMATİV ƏLAMƏTLƏRİN ÇIXARILMASI VƏ ÇƏKİLƏNMƏSİ METODUNUN ALQORİTMİK ƏSASLANDIRILMASI

Fərdi dil korpusları əsasında informativ əlamətlərin çıxarılması metodunun mahiyyəti yalnız mövcud məşin təlimi alqoritmlərinin tətbiqi deyil, müxtəlif linqvistik səviyyələrdən çıxarılan əlamətlərin informativlik indeksinə görə seçilməsi və vahid profil vektorunda birləşdirilməsidir. Təklif olunan yanaşma şərti olaraq 'Çoxsəviyyəli informativ əlamətlərin çıxarılması və çəkilənməsi metodu' adlandırılır. Metodun iş prinsipi Cədvəl 1-də mərhələli şəkildə verilmişdir.

Cədvəldən göründüyü kimi, mətnlər əvvəlcə fərdi korpusa çevrilir, emal və annotasiya olunur. Sonra müxtəlif linqvistik bloklardan əlamətlər çıxarılır, hər əlamət informativlik indeksinə görə çəkilənir və sonda ən güclü əlamətlər əsasında müəllif/profil vektoru yaradılır. Giriş verilənləri konkret müəllifə aid mətnlər çoxluğu $M = \{m_1, m_2, \dots, m_n\}$, mümkün olduqda janr, zaman, mənbə və region etiketləri, həmçinin dialektal, emosional və frazeoloji resurslardır. Çıxış nəticəsi isə müəllifi və ya müəllif profilini xarakterizə edən çoxsəviyyəli profil vektoru V və bu vektorda ən informativ əlamətlərin siyahısıdır.

A. Əlamət bloklarının formalaşdırılması

Hər bir mətn üçün əlamətlər səkkiz blok üzrə çıxarılır: L - leksik əlamətlər; M - morfoloji əlamətlər; S - sintaktik əlamətlər; D - diskurs-stilometrik əlamətlər; E - emosional əlamətlər; R - regional və dialektal əlamətlər; F - frazeoloji-mədəni əlamətlər; Q - yazı keyfiyyəti və oxunabilirlik əlamətləri. Beləliklə, hər bir mətn üçün ilkin əlamət vektoru aşağıdakı kimi verilir:

$$V_0 = [L, M, S, D, E, R, F, Q]$$

Bu strukturun üstünlüyü ondan ibarətdir ki, müəllif profili yalnız bir səviyyəli leksik göstəricilərə əsaslanmır, əksinə müəllifin dil davranışını leksik, qrammatik, emosional, regional və mədəni səviyyələrdə birlikdə təsvir edir.

CƏDVƏL I. ÇOXSƏVİYYƏLİ İNFORMATİV ƏLƏMƏTLƏRİN ÇIXARILMASI VƏ ÇƏKİLƏNMƏSİ METODUNUN MƏRHƏLƏLƏRİ

Mərhələ	Bu mərhələdə nə edilir?	Nəticə
1. Giriş verilənləri	Müəllifə aid mətnlər, metaverilənlər və linqvistik resurslar sistmə daxil edilir.	İlkin məlumat bazası
2. Korpusun formalaşdırılması	Mətnlər müəllif, zaman, janr, mənbə və mümkün olduqda region üzrə etiketlenir.	Fərdi dil korpusu
3. İlkin emal	Təmizləmə, tokenləşdirmə, normallaşdırma, lemmatizasiya və qeyri-standart yazılışların idarəsi aparılır.	Emal olunmuş mətnlər
4. Linqvistik annotasiya	POS, morfoloji, sintaktik, diskurs, emosional, dialektal və frazeoloji markerlər işarələnir.	Annotasiya olunmuş korpus
5. Əlamət blokları	Leksik, morfoloji, sintaktik, stilometrik, emosional, dialektal, frazeoloji və oxunabilirlik əlamətləri çıxarılır.	İlkin əlamət vektoru V_0
6. İnformativlik indeksi	Hər əlamət hədəf profil kateqoriyası ilə əlaqə, sabitlik, ayırdetmə gücü və interpretasiya qabiliyyətinə görə qiymətləndirilir.	Çəkilənmiş əlamətlər
7. Əlamətlərin seçilməsi	Aşağı informativlikli və təkrarlanan əlamətlər çıxarılır, ən güclü əlamətlər saxlanılır.	Seçilmiş əlamətlər F^*
8. Profil vektoru və identifikasiya	Seçilmiş əlamətlər vahid profil vektorunda birləşdirilir və klassifikasiya/identifikasiya mərhələsinə ötürülür.	Yekun profil vektoru V
9. Interpretasiya	Nəticəyə təsir edən əsas əlamətlər izah edilir: peşə yönümü, dialekt, emosional ton, dil sənəti və s.	İzah edilə bilən nəticə

B. İnformativlik indeksinin hesablanması

Təklif olunan metodda hər bir əlamətin identifikasiya dəyəri informativlik indeksi ilə qiymətləndirilir. Əlamət f üçün informativlik indeksi aşağıdakı ümumi model üzrə hesablanı bilər:

$$I(f) = \alpha_1 \cdot \text{Rel}(f, y) + \alpha_2 \cdot \text{Stab}(f) + \alpha_3 \cdot \text{Sep}(f) + \alpha_4 \cdot \text{Int}(f)$$

Burada $\text{Rel}(f, y)$ əlamətin hədəf profil kateqoriyası ilə statistik əlaqəsini, $\text{Stab}(f)$ əlamətin müxtəlif mətnlər və zaman kəsikləri üzrə sabitliyini, $\text{Sep}(f)$ fərqli müəllif və ya profil qruplarını ayırma qabiliyyətini, $\text{Int}(f)$ isə əlamətin linqvistik baxımdan izah edilə bilmə dəyərini göstərir. $\alpha_1, \alpha_2, \alpha_3$ və α_4 çəki əmsallarıdır və onların cəmi 1-ə bərabər qəbul edilir. Bu düstur metodun əsas alqoritmik yeniliyini təşkil edir: əlamətlər yalnız statistik gücünə görə deyil, həm də sabitlik və interpretasiya imkanına görə seçilir.

C. Profil vektorunun qurulması

Hər bir blok üzrə ən yüksək informativlik indeksinə malik əlamətlər seçilir və normallaşdırılır. Sonra blok çəkili tətbiq edilməklə yekun profil vektoru formalaşdırılır:

$$V = [\beta_L \cdot L^*, \beta_M \cdot M^*, \beta_S \cdot S^*, \beta_D \cdot D^*, \beta_E \cdot E^*, \beta_R \cdot R^*, \beta_F \cdot F^*, \beta_Q \cdot Q^*]$$

Burada $L^*, M^*, S^*, D^*, E^*, R^*, F^*$ və Q^* müvafiq bloklar üzrə seçilmiş əlamətlər, β isə hər blokun ümumi identifikasiya prosesində rolunu göstərən çəki əmsalıdır. Məsələn, regional mənsubiyyətin təyində β_R daha yüksək, yazı keyfiyyətinin təhlilində β_Q və β_S daha yüksək, peşə yönümünün müəyyənləşdirilməsində isə β_L daha yüksək götürülə bilər.

Alqoritmik təsvir: Çoxsəviyyəli informativ əlamətlərin çıxarılması:

Giriş: M - müəllifə aid mətnlər çoxluğu; R - linqvistik resurslar; y - hədəf profil kateqoriyası.

Çıxış: V - yekun profil vektoru; F^* - seçilmiş informativ əlamətlər.

Addım 1. M mətnlərini topla və mənbə, zaman, janr, region etiketləri ilə işarələ.

Addım 2. Hər mətn üçün təmizləmə, tokenləşdirmə, normallaşdırma və lemmatizasiya apar.

Addım 3. Mətnləri morfoloji, sintaktik, diskurs, emosional, dialektal və frazeoloji səviyyələrdə annotasiya et.

Addım 4. L, M, S, D, E, R, F və Q blokları üzrə ilkin əlamətləri çıxar.

Addım 5. Hər bir f əlaməti üçün $I(f)$ informativlik indeksini hesabla.

Addım 6. $I(f) \geq \tau$ şərtini ödəyən əlamətləri saxla, qalanlarını çıxar.

Addım 7. Seçilmiş əlamətləri blok çəkili ilə normallaşdır və V profil vektorunu qur.

Addım 8. V əsasında müəllif identifikasiyası, profil təsnifatı və ya regional oxşarlıq analizi apar.

VI. KLASSİFİKASIYA VƏ İDENTİFİKASIYA MƏRHƏLƏSİ

Yekun profil vektoru müxtəlif analitik məqsədlər üçün istifadə oluna bilər. Müəllif identifikasiyası üçün SVM, Loqistik regressiya (Logistic Regression), Təsadüfi meşə (Random Forest), Sadəlovh Bayes (Naive Bayes) kimi klassik alqoritmlər, həmçinin BiLSTM, transformer və ansambl modelləri tətbiq edilə bilər [7, 8, 16]. Regional mənsubiyyətin araşdırılmasında klasterləşdirmə, məsafə matrisləri və dialektometrik ölçülərdən istifadə etmək mümkündür [11, 15].

Bu mərhələdə əsas məqsəd yalnız yüksək dəqiqliyə nail olmaq deyil, həm də nəticənin izah oluna bilməsini təmin etməkdir. Məsələn, model müəllifi müəyyən regionla əlaqələndirərsə, bu nəticə hansı dialektal sözlər, morfosintaktik formalar və ya tezlik fərqləri əsasında alındığı göstərilməlidir. Eyni qayda emosional və frazeoloji profilləmə üçün də keçərlidir.

VII. MÜZAKİRƏ

Aparılmış təhlil göstərir ki, fərdi dil korpuslarının identifikasiyası üçün yalnız bir tip əlamətə söykənmək kifayət etmir. Leksik əlamətlər ilkin fərqləndirmə üçün səmərəli olsa da, müəllifin daha sabit və daha dərin xüsusiyyətlərini müəyyən etmək üçün sintaktik, diskurs, emosional, dialektal və frazeoloji səviyyələrin inteqrasiyası vacibdir [1, 2, 5]. Bu baxımdan çoxsəviyyəli linqvistik profil yanaşması daha əsaslı görünür.

Təklif olunan metod mövcud maşın təlimi alqoritmlərini sadalamaqla məhdudlaşmır. Onun fərqləndirici cəhəti əlamətlərin bloklar üzrə sistemləşdirilməsi, hər əlamət üçün informativlik indeksinin hesablanması, interpretasiya dəyərinin nəzərə alınması və profil vektorunun çəkili formada qurulmasıdır. Bu yanaşma xüsusilə Azərbaycan dili kimi resursları məhdud olan dillər üçün əhəmiyyətlidir, çünki az həcmli korpuslarda izah edilə bilən və linqvistik əsaslandırılmış əlamətlər daha etibarlı nəticə verə bilər.

Azərbaycan dili üçün əsas problemlərdən biri resurs çatışmazlığıdır. Fərdi korpusların etik və sistemli şəkildə toplanması, dialekt markerlərinin inventarlaşdırılması, emosional və frazeoloji resursların, morfosintaktik annotasiya sistemlərinin və etikətlənmiş təlim korpuslarının yaradılması vacibdir. Bununla yanaşı, mövcud beynəlxalq tədqiqatlar göstərir ki, zəruri korpus və alət bazası yaradıldığı halda Azərbaycan dili üçün çoxparametrlı fərdi korpus analizi və müəllif profiləməsi istiqamətində ciddi nəticələr əldə etmək mümkündür.

NƏTİCƏ

Tədqiqat işində fərdi dil korpusları əsasında identifikasiya üçün informativ əlamətlərin çıxarılması problemi kompleks şəkildə araşdırılmış və konkret mərhələlərə malik çoxsəviyyəli metod təklif edilmişdir. Metod korpusun formalaşdırılması, ilkin emal, annotasiya, əlamət bloklarının çıxarılması, informativlik indeksinin hesablanması, əlamətlərin seçilməsi, profil vektorunun qurulması və identifikasiya mərhələlərini birləşdirir.

Təklif olunan yanaşma müəllifin peşəkar yönümü, emosional vəziyyəti, dil səriştəsi, ləhcə-bölgə mənsubiyyəti, mədəni-folklor fonu, yazılı nitq keyfiyyəti və koqnitiv-lingvistik mürəkkəbliyi barədə nəticələr çıxarmağa imkan verən əlamətləri sistemləşdirir. Metodun əsas üstünlüyü statistik informativlik, sabitlik, ayırdetmə qabiliyyəti və interpretasiya dəyərini birlikdə nəzərə almasıdır. Bu yanaşmanın Azərbaycan dili üçün tətbiqi milli dil resurslarının inkişafına, regional dil xüsusiyyətlərinin avtomatik təhlilinə, sosial media analitikasına və izah edilə bilən süni intellekt sistemlərinin hazırlanmasına töhfə verə bilər.

ƏDƏBİYYAT

- [1] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Automatically profiling the author of an anonymous text," *Communications of the ACM*, vol. 52, no. 2, pp. 119–123, 2009.
- [2] D. Brunato, A. Cimino, F. Dell'Orletta, S. Montemagni, and G. Venturi, "Profiling-UD: A tool for linguistic profiling of texts," in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2020, pp. 7145–7151.
- [3] S. Ouni, F. Fkih, and M. N. Omri, "A survey of machine learning-based author profiling from texts analysis in social networks," *Multimedia Tools and Applications*, vol. 82, no. 24, pp. 36653–36686, 2023.
- [4] Y. HaCohen-Kerner, "Survey on profiling age and gender of text authors," *Expert Systems with Applications*, vol. 199, Art. no. 117140, 2022.
- [5] J. Soler-Company and L. Wanner, "On the relevance of syntactic and discourse features for author profiling and identification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2, 2017, pp. 681–687.
- [6] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches, "Overview of the author profiling task at PAN 2013," in *CEUR Workshop Proceedings*, vol. 1179, 2013.
- [7] J. A. García-Díaz, G. Beydoun, and R. Valencia-García, "Evaluating Transformers and Linguistic Features integration for Author Profiling tasks in Spanish," *Data and Knowledge Engineering*, vol. 151, Art. no. 102307, 2024.
- [8] J. Thomas, G. B. Hoang, and L. Mitchell, "Simple models are all you need: Ensembling stylometric, part-of-speech, and information-theoretic models for the ALTA 2024 Shared Task," in *Proceedings of the*

Australasian Language Technology Workshop, vol. 22, 2024, pp. 207–212.

- [9] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," *Engineering Reports*, vol. 2, no. 7, Art. no. e12189, 2020.
- [10] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Network Analysis and Mining*, vol. 11, no. 1, Art. no. 81, 2021.
- [11] B. Szmrecsanyi, "Corpus-based dialectometry: A methodological sketch," *Corpora*, vol. 6, no. 1, pp. 45–76, 2011.
- [12] P. Mishra and V. Mujadia, "Arabic dialect identification for travel and twitter text," in *ACL 2019 - 4th Arabic Natural Language Processing Workshop, WANLP 2019 - Proceedings of the Workshop*, 2019, pp. 234–238.
- [13] G. Andrioi and A. Moraru, "Cultural branding models in English and Romanian proverbs," *Cultural Perspectives*, vol. 24, pp. 27–44, 2019.
- [14] P. Ramírez Rodríguez, "The construction of cultural identity from phraseology: the case of Spanish proverbs by Russian SFL students," *Textos en Proceso*, vol. 9, no. 2, pp. 52–72, 2023.
- [15] D. Roemling, "Forensic authorship profiling using geolocated social media data: A corpus linguistic and cartographic approach," *Applied Corpus Linguistics*, vol. 5, no. 3, Art. no. 100146, 2025.
- [16] R. Alguliyev, R. Aliguliyev, and A. Gurbanova, "Method for Individuals' Speech Popularity Assessment Based on the National Terminological Information System," in *IEEE 17th International Conference on Application of Information and Communication Technologies (AICT)*, Baku, Azerbaijan, 2023, pp. 29–33, doi: 10.1109/AICT59525.2023.10313151.

Problems and Solutions of Multi-purpose Identification of Persons Based on Individual Language Corpora

Afruz Gurbanova¹, Sariya Gasimova²

¹Institute of Information Technology, Baku, Azerbaijan

²Nakhchivan State University, Nakhchivan, Azerbaijan

¹afruz1961@gmail.com, ²qasimovaseriyye@ndu.edu.az

Abstract- In the modern digital communication environment, individuals' written speech samples are becoming an important source for the creation of personal language corpora. This paper investigates the extraction of informative features for author identification and profiling based on individual language corpora. The main contribution of the study is a multi-level method that combines lexical, morphological, syntactic, discourse-stylometric, emotional, dialectal, phraseological-cultural, and readability-related indicators into a unified profile vector. The method evaluates the informativeness of each feature according to statistical relevance, stability, discriminative power, and interpretability. The proposed method includes corpus formation, preprocessing, linguistic annotation, feature extraction, informativeness index calculation, feature selection, and identification stages. The approach is promising for intelligent analysis of Azerbaijani personal language corpora, regional language variation detection, and explainable artificial intelligence applications.

Keywords- individual language corpus; author profile; informative features; dialectometry; emotional analysis; phraseological markers; readability.