Automated Linguistic Profiling of Disputed Texts in the Uzbek Language: A Model Based on Hybrid Vectorization and Support Vector Machine (SVM)

Mekhroj Raupov
raupovmehroj1994@gmail.com
Independent researcher,
Uzbekistan State University of World Languages

Annotation *The article proposes a model for the automated linguistic profiling of disputed texts in the Uzbek language. The aim is to develop a methodology that, for forensic-linguistic examination, determines in a quantitative, reproducible and interpretable manner the probabilistic socio-demographic characteristics of an author, namely gender, age and region, as well as the legal classification of a text into insult, defamation or neutral. The methodology integrates Biber's register analysis, Lakoff's language-and-gender theory and Nini's theory of linguistic individuality, adapting them to the agglutinative nature of Uzbek. Features are vectorized using a hybrid TF-IDF and FastText method, while a separate Support Vector Machine classifier is applied to each profiling task. The results are demonstrated through worked examples of TF-IDF weighting, character n-gram extraction and a confusion matrix. The proposed model operates interpretably and accurately under conditions of mixed Latin-Cyrillic writing and morphological richness. Thus, the study offers a codeable, interpretable and ethically constrained model for Uzbek forensic linguistics.*

Keywords *linguistic profiling, disputed text, forensic linguistics, support vector machine, hybrid vectorization, character n-grams, Uzbek language, idiolect*

O‘zbek tilidagi bahsli matnlarni avtomatlashtirilgan lingvistik profillash: gibrid vektorlashtirish va tayanch vektorlar mashinasi (SVM) asosidagi model

Raupov Mexroj Xukum o‘g‘li
raupovmehroj1994@gmail.com
Mustaqil izlanuvchi,
O‘zbekiston davlat jahon tillari universiteti

Annotatsiya *Maqolada o‘zbek tilidagi bahsli (nizoli) matnlarni avtomatlashtirilgan lingvistik profillash modeli taklif etiladi. Tadqiqotning maqsadi sud-lingvistik ekspertiza ehtiyojlari uchun matn muallifining ehtimoliy ijtimoiy-demografik tavsifini, ya‘ni jins, yosh va hududini, hamda matnning huquqiy tasnifini (haqorat, tuhmat yoki neytral) miqdoriy, takrorlanuvchan va izohlanuvchan tarzda aniqlovchi metodologiyani ishlab chiqishdan iborat. Metodologiya Biberning registr tahlili, Lakoffning til-jins nazariyasi va Ninining lisoniy individuallik nazariyasini o‘zbek tilining agglyutinativ tabiatiga moslashtirgan holda birlashtiradi. Belgilar gibrid usulda, ya‘ni TF-IDF va FastText vositalari yordamida vektorlashtiriladi; har bir profillash vazifasi uchun esa alohida tayanch vektorlar mashinasi tasniflagichi qo‘llaniladi. Tadqiqot natijalari TF-IDF og‘irlash, harf n-grammlarini ajratish va chalkashlik matritsasi misollarida amaliy ko‘rsatilgan. Taklif etilgan model o‘zbek tilidagi aralash lotin-kirill yozuv va morfologik boylik sharoitida ham izohlanuvchan,*

ham aniq ishlaydi. Shu tariqa tadqiqot o'zbek forensik lingvistikasi uchun kodlanadigan, izohlanuvchan va etik jihatdan chegaralangan model taqdim etadi hamda sud-ekspertiza amaliyotiga ilmiy asoslangan, milliy tilga moslashtirilgan yangi vosita olib kiradi. Model nazariy stilometriyani amaliy mashinaviy tasniflash bilan bog'laydi.

Kalit so'zlar *lingvistik profillash, bahsli matn, forensik lingvistika, tayanch vektorlar mashinasi, gibrid vektorlashtirish, harf n-grammlari, o'zbek tili, idiolekt*

Автоматизированное лингвистическое профилирование спорных текстов на узбекском языке: модель на основе гибридной векторизации и метода опорных векторов (SVM)

Раупов Мехрож Хукумович
raupovmehroj1994@gmail.com
Самостоятельный соискатель,
Узбекский государственный университет мировых языков

Аннотация *В статье предлагается модель автоматизированного лингвистического профилирования спорных текстов на узбекском языке. Цель исследования заключается в разработке методологии, которая для нужд судебно-лингвистической экспертизы количественно, воспроизводимо и интерпретируемо определяет вероятностную социально-демографическую характеристику автора, а именно пол, возраст и регион, а также правовую классификацию текста на оскорбление, клевету или нейтральный. Методология объединяет регистровый анализ Байбера, теорию языка и пола Лакофф и теорию лингвистической индивидуальности Нини, адаптируя их к агглютинативной природе узбекского языка. Признаки векторизуются гибридным методом, то есть с помощью инструментов TF-IDF и FastText; для каждой задачи профилирования применяется отдельный классификатор на основе метода опорных векторов. Результаты исследования продемонстрированы на примерах TF-IDF-взвешивания, выделения буквенных n-грамм и матрицы ошибок. Предложенная модель работает одновременно интерпретируемо и точно в условиях смешанного латинско-кириллического письма и морфологического богатства узбекского языка. Таким образом, исследование предлагает кодируемую, интерпретируемую и этически ограниченную модель для узбекской судебной лингвистики и вводит научно обоснованный, адаптированный к национальному языку инструмент в практику судебной экспертизы.*

Ключевые слова *Лингвистическое профилирование, спорный текст, судебная лингвистика, метод опорных векторов, гибридная векторизация, буквенные n-граммы, узбекский язык, идиолект*

Kirish

Raqamli kommunikatsiya hajmining keskin o‘sishi bilan ijtimoiy tarmoqlar, messenjerlar va izohlar maydonida yuzaga keladigan bahsli matnlar – haqorat, tuhmat, tahdid xarakteridagi yozma nutq namunalari – huquqiy va ijtimoiy muammoga aylandi. Bunday matnlarning muallifi ko‘pincha noma’lum yoki anonim bo‘lib, ularning kim tomonidan, qanday ijtimoiy-demografik guruh vakili tomonidan yozilganini aniqlash sud-lingvistik ekspertizaning dolzarb vazifasiga aylanmoqda. An’anaviy qo‘lda tahlil ham mehnattalab, ham ekspertning subyektiv mulohazasiga bog‘liq bo‘lib, katta hajmdagi raqamli materialni qamrab ololmaydi.

Mavzuning dolzarbligi milliy huquqiy kontekst bilan ham bog‘liq. O‘zbekiston Respublikasi qonunchiligida shaxsning sha’ni va qadr-qimmatini himoya qilish kafolatlangan: tuhmat va haqorat uchun jinoiy hamda ma'muriy javobgarlik nazarda tutilgan (O‘zbekiston Respublikasi Jinoyat kodeksi, 1994/2020). 2020-yilda kiritilgan o‘zgartishlarga ko‘ra, telekommunikatsiya yoki Internet tarmoqlarida joylashtirilgan haqorat va tuhmat materiallari uchun ham javobgarlik belgilandi. Bu – raqamli muhitdagi bahsli matnlar endi aniq huquqiy baholash obyektiga aylanganini ko‘rsatadi. Huquqiy ta'riflarni amalda qo‘llash esa matnning lingvistik tahlilini, ya’ni qaysi ibora tahqirlash yoki faktik uydirma, qaysi biri himoyalangan fikr ekanligini obyektiv ajratishni talab qiladi.

Tadqiqotning maqsadi – o‘zbek tilidagi bahsli matnlarni avtomatlashtirilgan lingvistik profillashning miqdoriy, takrorlanuvchan va izohlanuvchan modelini ishlab chiqish. Tadqiqotning ilmiy yangiligi: o‘zbek tili uchun moslashtirilgan gibrid (TF-IDF va FastText) belgilar fazosi taklif etildi; profillash va huquqiy tasnif vazifalari yagona belgilar vektoridan foydalanuvchi, ammo alohida o‘qitiladigan SVM tasniflagichlari sifatida ajratildi; modelning izohlanuvchanligi va etik chegaralari forensik kontekst talablariga muvofiq asoslandi. O‘zbek tilining agglyutinativ tabiati, aralash yozuvi va cheklangan til resurslari ingliz tili uchun yaratilgan modellarni to‘g‘ridan-to‘g‘ri qo‘llashga imkon bermaydi; shu bois milliy xususiyatlarni hisobga oluvchi maxsus model zarur.

Adabiyotlar tahlili

Matn uslubini miqdoriy o‘lchash g‘oyasi stilometriyaning klassik ishlariga borib taqaladi. Mosteller va Wallace anonim matnlarni funksional so‘zlar chastotasiga Bayes tahlilini qo‘llab muallifga nisbat berdilar (Mosteller & Wallace, 1964); ularning xulosasi – mazmunga bog‘liq bo‘lmagan funksional so‘zlar mualliflikning eng ishonchli ko‘rsatkichi ekanligi – bugungi stilometriyaning markaziy aksiomasi bo‘lib qolmoqda. Bu g‘oya keyinchalik standartlashtirilib, Delta o‘lchovi taklif etildi (Burrows, 2002). Registr variatsiyasini miqdoriy o‘rganishda leksik-grammatik belgilar faktor tahliliga tortilib, matnlar funksional o‘lchamlarga joylashtirildi (Biber, 1988).

Til va jins munosabati nazariyasi (Lakoff, 1975) keyinchalik hisoblash tilshunosligida empirik tasdiqlandi: muallif jinsini taxminan 80% aniqlikda, matnning badiiy yoki nobadiiyligini esa undan yuqori aniqlikda tasniflash mumkinligi ko‘rsatildi (Koppel va b., 2002). Bu yo‘nalish ijtimoiy tarmoq matnlariga kengaytirilib, katta hajmli korpusda jins va yoshni avtomatik profillash mumkinligi isbotlandi (Argamon, Koppel, Pennebaker, & Schler, 2009). Lisoniy individuallik nazariyasida mualliflikning eng samarali belgilari funksional so‘zlar chastotasi va kichik harf n-grammlari ekanligi asoslandi hamda forensik xulosalarni ehtimollik nisbati shaklida ifodalash taklif etildi (Nini, 2023).

Harf n-grammlarining kross-janr va kross-mavzu barqarorligi, shuningdek chastota belgilarni saralashning asosiy mezoni ekanligi ko‘rsatilgan (Stamatatos, 2009). Vektorlashtirish texnologiyasi vektor fazo

modelidan (Salton & McGill, 1983) atamaning o‘ziga xosligini statistik talqin qilish printsipigacha (Spärck Jones, 1972), undan subso‘z modeligacha rivojlandi; oxirgisi morfologik boy tillarda lug‘atdan tashqari so‘zlar muammosini hal qiladi (Bojanowski, Grave, Joulin, & Mikolov, 2017). Tasniflashda tayanch vektorlar mashinasi (Cortes & Vapnik, 1995) matn tasnifida samarali ekanligi empirik isbotlangan: u yuqori o‘lchovli, siyrak fazoga mos va ortiqcha moslashishga chidamli (Joachims, 1998). Matn tasnifi sohasidagi mashinaviy o‘rganish usullari keng umumlashtirilgan (Sebastiani, 2002), forensik kontekstda dalilni miqdoriy ifodalash esa alohida o‘rganilgan (Grant, 2007).

So‘nggi yillarda o‘zbek kompyuter lingvistikasida ham sezilarli natijalar to‘plandi. Milliy til korpusini yaratish, uning tuzilishi va imkoniyatlari tadqiq etildi (Elov & Alayev, 2023), korpus matnlarini raqamli shaklga o‘tkazishda esa TF-IDF, Word2Vec va BERT kabi vektorlashtirish usullarining o‘zbek tiliga tatbiqi o‘rganildi (Elov va b., 2023). Biroq bu ishlar asosan umumiy NLP vazifalariga (imlo tuzatish, so‘z turkumlarini teglash, mavzuviy tasnif) qaratilgan bo‘lib, sud-lingvistik profillash, ya'ni muallifning ehtimoliy demografik tavsifi va matnning huquqiy tasnifi, alohida tadqiqot predmeti sifatida o‘rganilmagan. Ushbu maqola aynan shu bo‘shliqni to‘ldiradi: u o‘zbek korpus lingvistikasining mavjud yutuqlariga tayanadi, ammo ularni xalqaro forensik stilometriya nazariyasi bilan birlashtirib, milliy sud-ekspertiza ehtiyojiga yo‘naltiradi. Shu ma’noda tadqiqot ikki yo‘nalish – o‘zbek hisoblash tilshunosligi va jahon forensik lingvistikasi – kesishmasida originallik kasb etadi.

Metodologiya

Model bosqichli quvur sifatida loyihalangan: xom matn, dastlabki ishlov, belgilar ajratish, vektorlashtirish, SVM bilan tasniflash va izohli chiqish. Dastlabki ishlovda matn normallashtiriladi, tokenlarga va morfemalarga ajratiladi; agglyutinativ tabiat hisobga olinib, so‘z o‘zak va affikslarga segmentlanadi. Aralash lotin-kirill yozuv yagona kodlashga keltiriladi, harf cho‘zish va emoji alohida belgilar sifatida qayd etiladi.

Belgilar fazosi uch qatlamdan tashkil topadi: uslubiy qatlam (registr o‘lchovi, ot/fe‘l nisbati, kuchaytiruvchilar zichligi) Biber (1988) asosida; demografik qatlam (yuklamalar, inkor, emoji va harf cho‘zish chastotasi) Lakoff (1975) va Koppel & Argamon (2002) asosida; individual qatlam (harf n-grammlari va funksional so‘zlar chastotasi) Nini (2023) va Burrows (2002) asosida. Bularga stilometrik o‘lchovlar (tip-token nisbati, o‘rtacha gap uzunligi) qo‘shiladi. Har bir qatlam profilning turli jihatini qamragani uchun yagona, lekin boy belgilar fazosi turli vazifalarga moslashuvchan asos beradi.

Vektorlashtirish gibrid usulda amalga oshiriladi. TF-IDF belgisi (Spärck Jones, 1972) belgining matndagi chastotasini uning korpusdagi nodirligi bilan muvozanatlaydi (1-formula):

$$TF\text{-}IDF(t, d) = tf(t, d) \times \log(N / df(t)) \quad (1)$$

Bu izohlanuvchan qatlam FastText subso‘z vektorlari (Bojanowski va b., 2017) bilan birlashtiriladi; birlashtirishdan oldin har bir guruh alohida normallashtiriladi. TF-IDF qaysi aniq belgi qarorga ta’sir qilganini ko‘rsatadi (forensik shaffoflik uchun zarur), FastText esa lug‘atdan tashqari so‘z shakllarini qamraydi. Tasniflashda har bir vazifa (jins, yosh, hudud, huquqiy tasnif) uchun alohida SVM o‘qitiladi; barchasi umumiy vektordan foydalanadi, ammo mustaqil baholanadi. Ko‘p sinfli vazifalar uchun “biri-barchaga qarshi” strategiyasi qo‘llaniladi, chiziqli yadro esa izohlanuvchanligi tufayli afzal ko‘riladi (Joachims, 1998). Model etik chegaralar bilan loyihalanadi: profil baholari ehtimoliy bo‘lib, alohida shaxsga deterministik tatbiq etilmaydi; kasb va psixologik portret kabi matndan ishonchli aniqlab bo‘lmaydigan tavsiflar asosiy natijalarga kiritilmaydi.

Belgilarni birlashtirish (feature fusion) bosqichi alohida ahamiyatga ega. Uch qatlam belgilari (uslubiy, demografik, individual) hamda FastText subso‘z vektorlari turli o‘lcham

va son qiymatiga ega bo‘lgani uchun, ularni to‘g‘ridan-to‘g‘ri ulash kattaroq qiymatli belgilarning boshqalarini bostirishiga olib keladi. Buning oldini olish uchun har bir guruh alohida normallashtiriladi (z-baho yoki L2-normallashtirish), so‘ngra yagona vektorga ketma-ket ulanadi (konkatenatsiya). Masalan, yuqorida ko‘rsatilgan bahsli gap uchun yakuniy vektor TF-IDF og‘irliklari (“bo‘lasan” = 0,1204 yetakchi), harf n-gramm chastotalari (inkor va shaxs affikslarini qamrovchi) hamda FastText vektorlaridan tashkil topadi. Shu tariqa lingvistik belgi sonli qiymatga, son esa huquqiy-lingvistik xulosaga aylanadi – bu jarayon modelning izohlanuvchanligini ta'minlaydi, chunki har bir bashorat ortida aniq, ko‘rsatib beriladigan lingvistik asos turadi.

Belgilarni amalda ajratish uchun model bir nechta maxsus lingvistik bazaga tayanadi, ular o‘zbek tilining va raqamli muhitning xususiyatlariga moslashtirilgan. Dialektologik baza hududiy sheva belgilarini, invektiv (haqoratli) leksika bazasi matnning huquqiy tasnifiga oid birliklarni, kuchaytiruv va ta'kid yuklamalari ro‘yxati esa gender va ekspressivlik markerlarini qamraydi. Alohida e'tibor grafemik bazaga qaratiladi: u aralash lotin-kirill yozuv, harf cho‘zish (“zo‘rrr”) va translit variantlarini qayd etadi, chunki bu hodisalar standart imloviy vositalar bilan qamralmaydi, ammo yosh guruhi va raqamli savodxonlik darajasining muhim ko‘rsatkichidir. Ushbu bazalarning mazmuni – aniq so‘z va qoidalar ro‘yxati – milliy til materiali asosida shakllantiriladi va modelni o‘zbek tiliga xos qiladi; aynan shu moslashuv tadqiqotning ilmiy hissasini tashkil etadi.

Natijalar

Modelning ishlashini bahsli gap misolida ko‘rsatamiz. Korpusda to‘rtta matn bo‘lib (N = 4), ulardan biri – “pulni qaytar yoki pushaymon bo‘lasan” (beshta so‘z shakli, har biri bir martadan, TF = 0,20). TF-IDF og‘irliklari 1-jadvalda keltirilgan.

So‘z shakli	df	IDF	TF-IDF
pulni	2	0,301	0,0602
qaytar	2	0,301	0,0602
yoki	2	0,301	0,0602
pushaymon	2	0,301	0,0602
bo‘lasan	1	0,602	**0,1204**

1-jadval. *Bahsli gap so‘zlarining TF-IDF og‘irliklari (N = 4)*

Natijadan ko‘rinadiki, faqat bahsli gapda uchragan “bo‘lasan” so‘zi eng yuqori og‘irlikka (0,1204) ega bo‘ldi – bu lingvistik jihatdan asosli, chunki aynan shu kelasi zamon shakli tahdidning shartli oqibatini ifodalovchi markaziy markerdir. Bu natija chastota belgilarni saralashning asosiy mezoni ekanligi haqidagi xulosaga (Stamatatos, 2009) mos keladi.

Harf n-grammlarini ajratish agglyutinativ morfologiyani lemmatizatsiyasiz qamraydi. Masalan, “kelmaysan” so‘zidan harf 3-grammlari ajratiladi: <ke, kel, elm, lma, may, ays, ysa, san, an>. Bu yerda may, ays, ysa, san trigrammlari inkor (-ma-) va ikkinchi shaxs (-san) affikslarini bevosita marker sifatida qayd etadi – bu o‘zbek tili uchun amaliy ustunlik (Nini, 2023).

O‘zbek tilining agglyutinativ tabiati belgilar ajratishda alohida e'tibor talab qiladi. Masalan, bitta “yoz-” o‘zagidan “yozdi”, “yozmadi”, “yozolmaysan”, “yozganlaridan” kabi o‘nlab so‘z shakli hosil bo‘ladi. Agar har bir shakl alohida belgi sifatida olinsa, vektor o‘lchami keskin kengayadi va ko‘pchilik belgi korpusda bir-ikki martagina uchrab, ishonchsiz signalga aylanadi. Aynan shu sababdan model so‘z shakllari o‘rnida harf n-grammlari va

subsoʻz (FastText) darajasida ishlaydi: bu yondashuv turli shakllarning umumiy tarkibiy qismlarini qamrab, ma'lumotlar siyrakligini kamaytiradi va morfologik jihatdan boy tilda barqaror natija beradi.

Teglangan korpus sifati modelning ishonchliligini bevosita belgilaydi. Shu bois huquqiy teglar (haqorat, tuhmat, neytral) bir nechta mustaqil ekspert tomonidan qoʻyiladi va ular oʻrtasidagi muvofiqlik annotatorlararo kelishuv (inter-annotator agreement) oʻlchovi, masalan Cohen kappa koeffitsiyenti yordamida baholanadi (Cohen, 1960). Yuqori kappa qiymati teglarning izchil va takrorlanuvchan ekanligini, demak oʻquv ma'lumotining ishonchli ekanligini koʻrsatadi; past qiymat esa teglash koʻrsatmalarini qayta koʻrib chiqish zarurligini bildiradi. Bunday nazorat forensik kontekstda ayniqsa muhim, chunki model xulosasi ekspert dalili sifatida ishlatilganda uning asosidagi ma'lumot sifati tekshiriluvchan boʻlishi shart.

Ikki matnning uslubiy yaqinligini miqdoriy oʻlchashda kosinus oʻxshashligi qoʻllaniladi. Masalan, ikki qisqa matn uchta belgi boʻyicha $d_1 = (2, 1, 0)$ va $d_2 = (1, 1, 1)$ vektorlari bilan ifodalansa, skalyar koʻpaytma 3 ga, vektor uzunliklari esa $\sqrt{5} \approx 2{,}236$ va $\sqrt{3} \approx 1{,}732$ ga teng boʻladi; demak kosinus oʻxshashligi $3 / (2{,}236 \times 1{,}732) \approx 0{,}775$. Birga yaqin qiymat ikki matnning uslubiy jihatdan oʻxshashligini koʻrsatadi. Matn uzunligi turlicha boʻlganda aynan burchak (yoʻnalish) oʻlchovi mutlaq masofadan ishonchliroq boʻladi, chunki u matn hajmiga bogʻliq emas.

Hudud profili oʻzbek tilining sheva xilma-xilligiga tayanadi. Dialektologik baza yordamida ayrim leksik va fonetik variantlar (masalan, Fargʻona, Xorazm yoki Qashqadaryo shevalariga xos soʻz shakllari) qayd etiladi va ularning chastotasi boʻyicha muallifning ehtimoliy hududi baholanadi. Bu belgilar boshqa qatlamlardan mustaqil boʻlib, profilga qoʻshimcha dalil qoʻshadi; biroq ular ham ehtimoliy boʻlib, alohida shaxsga qatʻiy bogʻlab boʻlmaydi, chunki bir muallif bir necha sheva belgilarini aralash qoʻllashi mumkin.

Modelning huquqiy tasnif vazifasi uch sinf (haqorat, tuhmat, neytral) boʻyicha baholandi (2-jadval).

Haqiqiy \ Bashorat	haqorat	tuhmat	neytral	Jami
haqorat	40	5	5	50
tuhmat	6	38	6	50
neytral	4	4	42	50
Jami	50	47	53	150

2-jadval. *Uch sinfli chalkashlik matritsasi (satr — haqiqiy, ustun — bashorat)*

Matritsadan har bir sinf uchun koʻrsatkichlar hisoblandi: “haqorat” — P = 0,80, R = 0,80, F1 = 0,80; “tuhmat” — P ≈ 0,809, R = 0,76, F1 ≈ 0,784; “neytral” — P ≈ 0,792, R = 0,84, F1 ≈ 0,815. Makro-oʻrtacha: P ≈ 0,80, R ≈ 0,80, F1 ≈ 0,80; umumiy aniqlik 120/150 = 0,80. (Keltirilgan qiymatlar hisob mantigʻini koʻrsatuvchi misol boʻlib, yakuniy natijalar muallif korpusida olinadi.)

Modelning amaliy ishonchliligini baholashda xatolik manbalarini tahlil qilish zarur. Oʻzbek tilidagi bahsli matnlarda eng qiyin holatlar quyidagilardir: inkor (masalan, “u ahmoq emas” – tashqi koʻrinishi haqoratli leksika, ammo ma'nosi inkor); kinoya va kesatiq (yuzaki neytral, ammo pragmatik jihatdan tahqirlovchi); hamda kod almashinuvi (oʻzbekcha-ruscha aralash nutq), bu belgilar fazosini chalgʻitishi mumkin. Bunday holatlar chalkashlik matritsasida xato musbat yoki xato manfiy sifatida namoyon boʻladi. Masalan, agar “tuhmat” koʻpincha “haqorat” deb xato tasniflansa, bu ikki sinfning leksik belgilari oʻxshashligini bildiradi va qoʻshimcha

farqlovchi belgi (ayblov konstruksiyalari yoki uchinchi shaxsga murojaat) talab qilinadi. Xato tahlili nafaqat modelni takomillashtiradi, balki ekspertga modelning cheklovlarini ochiq ko‘rsatib, xulosaning ishonchlilik darajasini to‘g‘ri baholashga yordam beradi.

Olingan natijalarning ishonchliligi

Ishonchlilik uch jihatdan asoslanadi. Birinchidan, metodologik asos: belgilar va algoritm jahon adabiyotida empirik tasdiqlangan natijalarga tayanadi (Koppel & Argamon, 2002; Nini, 2023; Joachims, 1998). Ikkinchidan, matematik-statistik nazorat: k-bo‘lakli kross-validatsiya qo‘llaniladi, sinflar nomutanosibligi tufayli asosiy mezon sifatida F1-o‘lchov tanlanadi, author-leakage oldini olish uchun bir muallif matnlari faqat bitta to‘plamga joylashtiriladi. Bundan tashqari, o‘quv korpusining teglash sifati annotatorlararo kelishuv (Cohen, 1960) bilan nazorat qilinadi, bu esa natijaning ma'lumot sifatiga bog‘liq qismini kafolatlaydi. Uchinchidan, qiyosiy asos: SVM bir xil fazoda boshqa tasniflagichlar bilan qiyoslanadi; nazariy jihatdan uning ustunligi kutiladi (Joachims, 1998; Sebastiani, 2002), biroq bu xulosa muallifning empirik natijalari bilan tasdiqlanishi shart. Forensik kontekstda model qat'iy hukm emas, ehtimollik nisbati shaklidagi dalil beradi (Grant, 2007); yakuniy xulosa malakali ekspert zimmasida qoladi.

Taklif etilgan modelning amaliy ahamiyati uning hisoblash tejamkorligi bilan ham bog‘liq. O‘qitilgandan so‘ng tayanch vektorlar mashinasi yangi matnni juda tez tasniflaydi va minglab xabar, izoh hamda postdan iborat katta hajmli materialni qisqa vaqtda qayta ishlay oladi; bu sud-ekspertiza amaliyotida muhim, chunki ekspert bunday ko‘lamdagi materialni qo‘lda tahlil qila olmaydi. Model nisbatan kam hisoblash resursi talab qilgani uchun maxsus apparatsiz, oddiy ish stansiyasida ham ishlaydi va milliy ekspertiza muassasalarida joriy etishga qulay. Shu tariqa nazariy stilometriya yutuqlari amaliy, kodlanadigan va tashkiliy jihatdan bajariluvchi vositaga aylanadi.

Xulosa

Tadqiqotda o‘zbek tilidagi bahsli matnlarni avtomatlashtirilgan lingvistik profillashning yaxlit modeli taklif etildi. Model xalqaro olimlarning empirik asoslangan metodologiyalarini (Biber, 1988; Lakoff, 1975; Nini, 2023; Koppel & Argamon, 2002) o‘zbek tilining agglyutinativ tabiati va aralash raqamli yozuv muhitiga moslashtirib, gibrid (TF-IDF va FastText) vektorlashtirish hamda alohida SVM tasniflagichlari asosida ishlaydi. Natijalar amaliy misollarda ko‘rsatildi; izohlanuvchanlik va etik chegaralar forensik talablarga muvofiq asoslandi. Amaliy ahamiyati – model sud-lingvistik ekspertiza uchun miqdoriy, takrorlanuvchan va kodlanadigan asos beradi. Kelajakdagi yo‘nalishlar: teglangan korpusni kengaytirish; izohlanuvchan SVM belgilarini kontekstual modellar bilan uyg‘unlashtiruvchi gibrid ansambl yaratish; modelni real ekspertiza ish oqimiga joriy etish.

References:

1. Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM, 52*(2), 119–123.
2. Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.
3. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5,* 135–146.
4. Burrows, J. (2002). ‘Delta’: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267–287.

5. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20(*1), 37–46.
6. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.
7. Elov, B., & Alayev, R. (2023). O‘zbek tili korpusi va uning imkoniyatlari. *O‘zbekiston informatika va energetika muammolari jurnali,* (2).
8. Elov, B., Hamroyeva, Sh., Alayev, R., Xusainova, Z., & Yodgorov, U. (2023). O‘zbek tili korpusi matnlarini qayta ishlash usullari. *Raqamli transformatsiya va sun’iy intellekt, 1*(3), 117–130.
9. Grant, T. (2007). Quantifying evidence in forensic authorship analysis. *International Journal of Speech, Language and the Law, 14*(1), 1–25.
10. Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features.* In Proceedings of ECML-98 (LNCS 1398, pp. 137–142). Springer.
11. Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing, 17*(4), 401–412.
12. Lakoff, R. (1975). *Language and woman’s place.* Harper & Row.
13. Mosteller, F., & Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist.* Addison-Wesley.
14. Nini, A. (2023). *A theory of linguistic individuality for authorship analysis.* Cambridge University Press.
15. O‘zbekiston Respublikasi Jinoyat kodeksi. (1994/2020). *139–140-moddalar.* Toshkent. https://lex.uz/docs/-111453
16. Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* McGraw-Hill.
17. Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.
18. Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.
19. Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538–556.