

e-ISSN 3060-5059

DOI: 10.67227



IJTIMOIIY-GUMANITAR SOHADA ILMIY-INNOVATSION TADQIQOTLAR

ILMIY METODIK JURNAL

Scientific and Innovative Research in
Social and Humanitarian Studies

Vol. 3 • Issue 10

2026



Founded in 2024



publishscience.uz



publish_science_journal@mail.ru



PEER REVIEWED • OPEN ACCESS

O'ZBEK TILIDAGI BAHSLI MATNLARNI AVTOMATIK TASNIFLASH VA MUALLIFNI LINGVISTIK PROFILLASH

Raupov Mexroj Xukum o'g'li

O'zbekiston davlat jahon tillari universiteti, mustaqil tadqiqotchi.

e-mail: raupovmehroj1994@gmail.com ORCID: <https://orcid.org/0009-0000-5502-0146>

Annotatsiya.

Maqolada o'zbek tilidagi bahsli matnlarni avtomatik tasniflash va muallifning ehtimoliy profilini aniqlash masalalari tadqiq etiladi. 400 birlikdan iborat sintetik UCTS korpusida izohlanuvchan model muzlatilgan test to'plamida baholandi: kontekst toifalari bo'yicha makro-F1 ko'rsatkichi 0,712 ni tashkil etdi. Muallif jinsini aniqlash natijasi nisbatan past (0,567) bo'lib, bu ko'rsatkich qat'iy dalil emas, balki yo'naltiruvchi indikator sifatida qaralishi lozimligini ko'rsatadi.

Kalit so'zlar: bahsli matn, avtomatik tasniflash, UCTS korpusi, sud-lingvistik ekspertiza, lingvistik profilash, makro-F1, annotatorlararo kelishuv, forensik lingvistika.

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ СПОРНЫХ ТЕКСТОВ НА УЗБЕКСКОМ ЯЗЫКЕ И ЛИНГВИСТИЧЕСКОЕ ПРОФИЛИРОВАНИЕ АВТОРА

Раупов Мехрож Хукум угли

Узбекский государственный университет мировых языков, независимый исследователь.

Аннотация.

В статье исследуются вопросы автоматической классификации спорных текстов на узбекском языке и вероятностного профилирования их авторов. Интерпретируемая модель была оценена на фиксированной тестовой выборке синтетического корпуса UCTS, состоящего из 400 единиц. Показатель макро-F1 по категориям контекста составил 0,712. Результат определения пола автора оказался сравнительно низким (0,567), что свидетельствует о необходимости рассматривать данный показатель не как достоверное доказательство, а исключительно как ориентировочный индикатор.

Ключевые слова: спорный текст, автоматическая классификация, корпус UCTS, судебно-лингвистическая экспертиза, лингвистическое профилирование, макро-F1, межаннотаторское согласие, форензическая лингвистика.

AUTOMATIC CLASSIFICATION OF CONTROVERSIAL TEXTS IN UZBEK AND LINGUISTIC PROFILING OF THE AUTHOR

Raupov Mexroj Xukum o'g'li

Uzbekistan State World Languages University, Independent Researcher.

Abstract.

The article investigates the automatic classification of disputed texts in Uzbek and the probabilistic profiling of their authors. An interpretable model was evaluated on a fixed test set of the synthetic UCTS corpus comprising 400 units. The macro-F1 score across contextual categories reached 0.712. The performance of author gender prediction was relatively low (0.567), indicating that this measure should be regarded not as conclusive evidence but merely as an indicative metric.

Keywords: disputed text, automatic classification, UCTS corpus, forensic linguistic examination, linguistic profiling, macro-F1, inter-annotator agreement, forensic linguistics.

Kirish. Raqamli muloqot makonining jadal rivojlanishi kommunikatsiyani ommaviy, tezkor va ko'p hollarda anonim ko'rinishga keltirdi. Ijtimoiy tarmoqlar va messenjerlarda muallifi noma'lum yoxud soxta akkaunt ortida yashiringan, haqorat, tahdid, tuhmat va ijtimoiy nizo qo'zg'atish mazmunidagi bahsli matnlar hajmi ortib bormoqda. Bunday matnlarni qo'lda tahlil qilish ularning ko'lami va tezligidan ortda qolmoqda; shu bois takrorlanuvchan va tekshiriluvchan avtomatik tahlil vositalariga ehtiyoj kuchaymoqda. Mamlakatimizda davlat xizmatlarini raqamlashtirish va murojaatlarni avtomatik qayta ishlash vazifalari ham o'zbek tilida "fikrlovchi" matn tahlili dasturlarini yaratish zaruratini keltirib chiqaradi.

Mavzuning dolzarbligi ikki jihat bilan belgilanadi. Nazariy jihatdan — sud-lingvistik ekspertiza va tabiiy tilni qayta ishlash kesishmasida o'zbek tili uchun yangi, izohlanuvchan modellarni asoslash zarur; amaliy jihatdan — anonim va soxta akkauntlar ortidagi shaxsni aniqlash, murojaatlarni avtomatik saralash va kiberxavfsizlik vazifalari takrorlanuvchan, miqdoriy tahlil vositalarini talab qiladi. Shu bilan birga, avtomatik tizim ekspertning o'rnini bosmaydi, balki unga dastlabki, miqdoriy asoslangan dalil taqdim etadi; yakuniy huquqiy-lingvistik xulosa malakali mutaxassis

zimmiasida qoladi (inson nazorati tamoyili).

Tadqiqotning maqsadi — o'zbek tilidagi bahsli matnlarning maxsus korpusini (Uzbek Controversial Texts Subcorpus, UCTS) shakllantirish, uni izohlanuvchan tasniflash modeli bilan baholash hamda muallifning ehtimoliy demografik profilini aniqlash imkoniyatlarini tekshirishdan iborat. Dastur ketma-ket to'rtta masalani yechadi: bahslilik toifasi (T1, to'rt sinf: haqorat, tuhmat, tahdid, ijtimoiy nizo), muallif jinsi (T2), yosh guruhi (T3) va platforma/registr (T4). Tadqiqot savollari quyidagilardan iborat: izohlanuvchan model bahsli matnlarni toifalarga qanchalik aniq ajratadi; pragmatik qoidalar leksik markerlarga nisbatan qanday hissa qo'shadi; muallif jinsini matn asosida ehtimoliy baholash qanchalik ishonchli?

ADABIYOTLAR TAHLILI VA METODOLOGIYA

Sud (forensik) lingvistikasining nazariy o'zagini matnni dalil sifatida talqin qilish g'oyasi tashkil etadi [1: 7]. Har bir shaxs nutqining betakrorligi — idiolekt — matnni muallifga bog'lovchi asos bo'lib, uning daliliy qiymati farqlovchilik va izchillik mezonlariga bo'ysunadi; lisoniy belgi muallifni "yuz foiz" emas, balki ma'lum ehtimollik bilan ko'rsatadi [14: 12]. Stilometriyaning rivojlanishi intuitiv uslubiy mulohazalarni miqdoriy ko'rsatkichlarga — funksional so'zlar chastotasi, leksik xilma-xillik, n-gramm profiliga aylantirdi [13; 12]. Xalqaro NLP tadqiqotlarida nafrat nutqi va toksik matnlarni avtomatik aniqlash alohida yo'nalishga aylanib, samaradorlikning kaliti lingvistik jihatdan asoslangan belgilar tanlovi va sifatli etiketlangan korpus ekanligi ko'rsatildi [10].

Mazkur tadqiqotning nazariy negizi uchta fundamental yo'nalishni yagona belgi fazosiga birlashtiradi: D. Biberning ko'p o'lchovli registr tahlili matnning uslubiy profilini [5], R. Lakoffning til-gender nazariyasi demografik profilni [6], A. Ninining n-gramm stilometriyasi esa individual mualliflik profilini aniqlashga xizmat qiladi [13]. Korpus lingvistikasida representativlik tanlov makonini aniq belgilash va undan tizimli namuna olish orqali ta'minlanadi [5: 24], annotatsiyaning ishonchligi esa kelishuv koeffitsiyenti bilan baholanadi: J. Landis va G. Koch talqinicha, κ ning 0,61–0,80 oralig'i "sezilarli", 0,81 dan yuqorisi esa "deyarli to'liq" kelishuvni bildiradi [24: 165].

O'zbek tilshunosligida profilashning nazariy asoslari uslubshunoslik, pragmalingsvistika va leksikologiya sohalarida shakllangan: Sh. Safarovning pragmalingsvistik qarashlari muallif niyati va nutqiy strategiyasini aniqlashda mustahkam poydevor beradi [16]. Biroq o'zbek tilining agglutinatib tabiati profilashga o'ziga xoslik kiritadi: morfologik ko'rsatkichlarning o'ta mahsuldorligi muallif idiolektining boy manbai bo'lsa-da, so'z shakllarining ko'pligi ma'lumotlar siyrakligini keltirib chiqaradi. Shu bois so'z darajasidagi belgilardan belgi (subso'z) n-grammlariga o'tish morfologik boy tillar uchun afzal yondashuv sifatida tanlandi [9]. Mavjud umumiy o'zbek korpuslari asosan adabiy va neytral matnlarga yo'naltirilgan bo'lib, bahsli matnlarga ixtisoslashgan, metama'lumotli etalon baza ishlab chiqilmagan — bu esa ushbu tadqiqotning ilmiy bo'shlig'ini belgilaydi.

Korpus.

Tadqiqot materiali — UCTS korpusining sinov versiyasi: 400 ta matn birligi, jami 8 231 ta so'z (o'rtacha 20,6 ta so'z), XSD sxemasi bo'yicha tasdiqlangan, takrorsiz va yagona identifikatorga ega. Birliklar to'rtta kontekst toifasi bo'yicha stratifikatsiyalangan (ijtimoiy nizo — 120, tahdid — 120, haqorat — 80, tuhmat — 80), platforma, muallifning shartli jinsi (erkak — 206, ayol — 194) va yosh guruhi bo'yicha muvozanatlangan. Korpus to'liq sintetik bo'lib, real shaxs, tashkilot, ism, manzil yoki aloqa ma'lumotlarini saqlamaydi; bu sud-lingvistik tadqiqotlarning etik talablariga muvofiqdir.

Dastlabki ishlash va leksik baza.

Matnlar yagona normallashtirish moduli orqali qayta ishlandi (kirill → lotin, "e → ye", tutuq belgisi variantlarini birxillashtirish), so'ng tokenizatsiya va lemmatizatsiyadan o'tkazildi; o'zbek tiliga xos morfologik hodisalar (apostrof variantlari, $h \leftrightarrow x$ va $q \leftrightarrow k$ almashinuvi, "moq" bilan tugovchi so'zning fe'l/ot farqi) maxsus qoidalar bilan hal etildi. Markerlarni aniqlash uchun 6 635 birlikdan (faol — 4 826) iborat yagona reyestr qo'llanildi: dialektal — 2 040, invektiv — 1 778, funksional — 596, tuhmat — 538, tahdid — 493, jargon — 420, murojaat — 333, nizo — 182, agressiv — 107, hissiy — 55. Har birining teglar to'plami, forensik kuchi (0–3) va kontekst sharti "O'zbek tilining izohli lug'ati"ning ikkala nashri [25; 26] asosida jild va bet ko'rsatkichlari bilan tekshirildi.

O'zbek tilida lemmatizatsiya alohida murakkablik tug'diradi: bir o'zakka o'nlab affikslar ketma-ket ulanadi, affikslar omonimiyasi va tovush o'zgarishlari esa bir so'zshaklning bir necha xil tahlilga ega bo'lishiga sabab bo'ladi. Masalan, "olma" so'zi ham ot, ham fe'lning bo'lishsiz shakli bo'lishi mumkin; to'g'ri lemma faqat kontekst asosida aniqlanadi. Shu bois soddastemming yetarli emas, cheklangan holatli avtomat (finite-state) asosidagi morfologik tahlil zarur bo'ladi [17: 44]. Mazkur tadqiqotda "moq" bilan tugovchi so'zning fe'l yoki ot ekanligi forensik jihatdan muhim qoida bilan hal etildi: bunday so'z sukut bo'yicha fe'l deb olinadi, ot sifatida esa faqat yopiq ro'yxatda (15 ta tub va 18 ta qo'shma ot — ahmoq, qamoq, muzqaymoq, beshbarmoq kabi) bo'lsa, qaraladi. Bu qoida tufayli "ko'tarmoq", "yuksalmoq" fe'l, "muzqaymoq" esa ot sifatida to'g'ri ajratildi.

Mualliflik atributsiyasi va profilash jahon tilshunosligida empirik baholashni talab qiladi: muallif identifikatsiyasi usullarining ishonchligi nazorat ostidagi eksperimentlar orqali tasdiqlanishi lozim [3: 540; 14: 12]. Ayniqsa, harf (belgi) n-grammlari mualliflik atributsiyasining eng barqaror belgilaridan biri sifatida e'tirof etilgan: ular bir vaqtning o'zida leksik, morfologik va imloviy odatlarni qamrab oladi hamda lug'atdan tashqari so'zlarga chidamli bo'ladi [13]. O'zbek tili uchun bu ayniqsa qimmatli, chunki affiks ketma-ketliklari harf n-grammlarida to'liq lemmatizatsiyasiz qayd etiladi; shu sababli mazkur tadqiqotda belgi 3–5-grammlari Nini qatlamining asosiy belgisi sifatida tanlandi.

Belgi fazosi.

Belgi vektori yetti qatlamdan iborat: (A) leksik-pragmatik markerlar; (B) Biber registr belgilari [5]; (C) Lakoff sotsiolingsvistik belgilari [6; 11]; (D) Nini n-grammlari [13; 12]; (E) kasbiy muhit (11 797 birlik, 23 soha); (F) psixolingsvistik qatlam; (G) muallif barqarorligi qatlami. Registr qatlamida leksik xilma-xillik Type-Token Ratio va undan barqarorroq bo'lgan harakatlanuvchi oynali ko'rsatkich orqali o'lchanadi:

$$TTR = V / N; MATTR(w) = (1/(N-w+1)) \cdot \sum TTR_i,$$

bu yerda V — turli lemmalar soni, N — jami soʻzlar soni, w — oyna uzunligi. N -gramm qatlamida belgi 3–5-grammlari va soʻz 1–2-grammlari TF-IDF usulida, sublineyar silliqlash bilan vaznlandi [21; 22]:

$$tf-idf(t, d) = (1 + \log-freq(t, d)) \cdot (\log[(1 + nd)/(1 + df(t))]) + 1.$$

Qatlamlar yagona vektorga birlashtirildi (FeatureUnion) va faqat oʻquv qismida moslandi — bu test maʼlumotlarining sizib chiqishini (data leakage) oldini oladi; forensik “kuch” (eng yuqori qiymat) va kontekst sharti birlashtirishda saqlandi.

Qatlamlarning batafsil tavsifi.

Sotsiolingvistik (Lakoff) qatlam 777 birlik va 16 sinfdan iborat psixolingvistik leksikonga (hissiy soʻzlar — 302, haqorat — 66, tilak-duo — 66, undov — 61, modal-qatʼiy — 34, kuchaytiruvchi — 30 va boshqalar), 90 ta emoji-hissiyot jadvaliga va grafemik belgilarga (emoji zichligi, harflarni choʻzish, tinish belgilarining takrorlanishi, KATTA HARFLAR ulushi) tayanadi. Kasbiy muhit qatlami 23 sohani qamrab oluvchi 11 797 birlikdan iborat leksikonga asoslanadi; leksikon manbasi — “Oʻzbek tilining izohli lugʻati”ning olti jildli nashri boʻlib, soha belgilari sahifa koordinatalari boʻyicha optik belgilarni tanish (OCR) orqali qayta oʻqildi. Psixolingvistik qatlamda jins modeli 13 ta belgidan (1-shaxs olishi, “siz”, hissiy-baho, xavotir-qoʻrquv, ijobiy hissiyot, kuchaytiruvchi, qatʼiylik, xushmuomalalik, murojaat, tilak-duo, uzun soʻz, uzun gap) iborat boʻlib, Naive Bayes log-ehtimollik nisbati va logistik kalibrlash bilan ishlaydi.

Qatlamning markaziy forensik qoidasi — hissiyot va agressiyaning qaratilganlik bilan darvozalanishi: marker faqat shaxsga (2-shaxs yoki aniq nishon) qaratilganda sanaladi, neytral 3-shaxs gapi qoʻshni markerdan hissiyotni “meros” qilib olmaydi. Masalan, “Mol goʻshti qimmatladi” yoki “Advokat sudda nutq oʻqidi” gaplari neytral deb baholanadi. Muallif barqarorligi qatlami esa grafemik va imloviy belgilardan faqat muallif barqarorligi DEVda tasdiqlangan (sinflararo korrelatsiya va AUC boʻyicha) belgilarni — gapning oʻrtacha uzunligi, kirill harflari ulushi, imloviy ogʻishlar ulushi kabilarni — muallifni ajratuvchi belgi sifatida qoldiradi; tahrir taʼsiriga uchraydigan belgilar faqat tavsifiy maqomda saqlanadi.

Koʻp vazifali birliklardagi ziddiyatlar forensik tamoyil asosida hal qilindi: bir soʻz bir nechta tegga ega boʻlsa, teglar toʻplami saqlanadi (29 holat); bir birlikning bir nechta kuch darajasi boʻlsa, eng yuqorisi olinadi (22 holat, hech qachon kamaytirilmaydi); faol qaratilgan haqorat, tuhmat yoki tahdid ijobiy yoki kasbiy talqinni bosib ketadi (5 holat), qaratilgan soʻkish kasbiy sohaga kiritilmaydi (106 ta ustma-ustlik). Korpus, shuningdek, koʻp qatlamli teglash bilan boyitildi: metamaʼlumotlar, leksik marker teglari (invektiv, tuhmat, tahdid) va nutqiy akt (pragmatik) teglari. Pragmatik teglash markaziy ahamiyatga ega, chunki haqorat, tahdid va tuhmat illokutiv kuch masalasidir: haqorat ekspressiv, tahdid komissiv-direktiv, tuhmat esa yolgʻon assertiv aktdir [15]; bir xil leksik birlik (“koʻrasan”) neytral kontekstda bashorat, agressiv kontekstda esa tahdid sifatida teglanadi.

Tasniflash.

Asosiy tasniflagich — tayanch vektorlar mashinasi (SVM) [18], qaror funksiyasi:

$$f(x) = \text{sign}(\sum \alpha_i y_i K(x_i, x) + b), \quad K(x_i, x) = \exp(-\gamma \|x_i - x\|^2).$$

Ehtimollik baholari Platt silliqlashi bilan kalibrlanadi [19]; ishonch 0,6 dan past boʻlsa, natija “aniqlab boʻlmadi” deb beriladi. Qiyoslash uchun bazaviy modellar (Naive Bayes, logistik regressiya, tasodifiy oʻrmon, kNN, Dummy) qoʻllanadi; giperparametrlar ichma-ich kross-validatsiyada tanlanadi [20], butun jarayon bitta boshlangʻich qiymat (seed 42) bilan takrorlanuvchan boʻladi.

Annotatsiya va sunʼiy intellektdan foydalanish (oshkorlik).

Jurnal talabiga muvofiq, sunʼiy intellektdan foydalanish oshkor etiladi: korpus annotatsiyasi va leksik reyestr har bir maydon boʻyicha ikkita mustaqil til modeli annotatori (Claude Opus 5.5 va Claude Sonnet 5.5) hamda kelishmovchiliklarni hal qiluvchi hakam-model protokoli yordamida shakllantirilgan. Barcha annotatsiya qarorlari avtomatik boʻlib, inson-ekspert validatsiyasi keyingi bosqich vazifasi sifatida saqlanadi. Annotatorlararo kelishuv Koen kappasi [23] va bootstrap 95% ishonch oraligʻi [29] orqali, tartibli maydonlar uchun esa vaznli kappa bilan [24] oʻlchandi. Oltin toʻplam sifati (annotatorlararo κ) va detektor aniqligi doimo alohida koʻrsatkichlar sifatida keltiriladi.

Baholash.

Asosiy metrika — makro-F1 (sinflar nomutanosib boʻlgani uchun); qoʻshimcha metrikalar — aniqlik, vaznli F1 va har bir sinf boʻyicha $P/R/F1$:

$$P = TP/(TP+FP), \quad R = TP/(TP+FN), \quad F1 = 2PR/(P+R), \quad \text{makro-F1} = (1/|Y|) \cdot \sum F1_c.$$

Korpus stratifikatsiyalangan DEV (280) va TEST (120) toʻplamlariga ajratildi (seed 42); sozlash faqat DEVda amalga oshirildi, muzlatilgan TEST esa yakunda bir marta baholandi. Farqlarning ahamiyati McNemar/Wilcoxon mezonlari, markerlarning toifalar kesimidagi bogʻliqligi χ^2 va Kramer V bilan, koʻp taqqoslashlar uchun Holm tuzatishi bilan tekshirildi.

Eksperiment dizayni.

Model samaradorligi va alohida belgi qatlamlarining hissasi ketma-ket tajribalar (E1–E8) orqali baholandi: E1/E2 — kontekst toifalarini tasniflash; E2b — kontekst-qoida ablatsiyasi; E7 — muallif profilini (jins, yosh) baholash; qolgan tajribalar alohida belgi qatlamlarining (imlo, frazeologiya, psixolingvistik) mustaqil detektor sifatidagi aniqligini oʻlchaydi. Bazaviy modellar bilan qiyoslashda koʻpchilik sinf (Dummy) tasniflagichi va faqat leksikonga asoslangan tasniflagich mezon sifatida olindi; marker va qoida qatlami ularning ikkalasidan ham ustun chiqdi. Barcha tajribalar faqat DEVda sozlanib, held-out TESTda bir marta baholandi — bu ortiqcha moslashuv va test maʼlumotlarining sizib chiqishini oldini oluvchi asosiy shartdir.

NATIJALAR

Bahslilik toifasini tasniflash muzlatilgan held-out testda ($n = 120$) makro-F1 0,712 va umumiy aniqlik 0,717 natijasini berdi; DEV to'plamidagi ko'rsatkich (0,733) bilan farq kichik, ya'ni ortiqcha moslashuv kuzatilmadi. Toifalar kesimidagi natijalar 1-jadvalda keltirilgan.

1-jadval. Kontekst toifalari bo'yicha tasniflash natijalari (held-out TEST, $n=120$)

Toifa	F1	Izoh
Haqorat	0,76	aniq invektiv leksika
Ijtimoiy nizo	0,74	jamoaviy/umumlashtiruvchi birliklar
Tuhmat	0,73	faktiv ayblovl konstruksiyalari
Tahdid	0,61	P 1,00 / R 0,44 — yashirin/shartli tahdid
Makro-o'rtacha	0,712	aniqlik (accuracy) 0,717

Pragmatik qoidalarining hissasini aniqlash uchun ablatsiya tajribasi o'tkazildi: faqat leksik markerlar bilan makro-F1 ko'rsatkichi 0,641 ni tashkil etdi; pragmatik qoidalar (ikkinchi shaxsga qaratilgan kamsitishni haqoratga, jamoaviy qarama-qarshilikni ijtimoiy nizoga yo'naltirish) qo'shilganda esa 0,72 ga ko'tarildi. Demak, bu qoidalarni olib tashlash F1 ko'rsatkichini 0,135 ballga pasaytiradi. Bu natija izohlanuvchan pragmatik qoidalarining yakka markerlarni sanashdan ustunligini ko'rsatadi.

Belgi qatlamlarining umumiy natijaga hissasi notekis bo'ldi. Leksik-pragmatik markerlar va pragmatik qoidalar (A qatlami) kontekst toifasini aniqlashda yetakchi rol o'ynadi; Biber va Nini qatlamlari (B, D) uslubiy va individual kontekstni to'ldirdi; psixolingvistik va muallif barqarorligi qatlamlari (F, G) esa asosan demografik baholarga xizmat qildi. Ablatsiya shuni ko'rsatdiki, eng katta yakka hissani pragmatik qoidalar beradi (0,135 ball): bu izohlanuvchan, qoidaga asoslangan qatlamning yalpi markerlarni sanashdan ustunligini yana bir bor tasdiqlaydi va modelning forensik kontekstdagi shaffoqligini ta'minlaydi.

Annotatsiya qatlamlarining ichki sifati (annotatorlararo kelishuv) yuqori bo'ldi (2-jadval). Ta'kidlash lozimki, bu qiymatlar oltin to'plam sifatini ifodalaydi va detektor aniqligi bilan aralashtirilmaydi: hissiyot bo'yicha annotatorlararo κ 0,911 bo'lsa-da, detektorning oltin to'plam bilan haqiqiy mosligi ancha kamtarona (κ 0,364).

Eslatma: Birinchi xatboshidagi raqamlar o'zaro mos emas: $0,720 - 0,641 = 0,079$, 0,135 emas. Agar boshlang'ich va yakuniy F1 qiymatlari to'g'ri bo'lsa, ikkala joydagi 0,135 qiymatini 0,079 ga almashtirish kerak.

2-jadval. Annotatsiya qatlamlari bo'yicha annotatorlararo kelishuv (Koen κ)

Qatlam / maydon	κ
Bahslilik leksikasi — teg	0,94
Bahslilik leksikasi — kuch (vaznli)	0,76
Psixolingvistik leksikon — nutqiy akt	0,946
Psixolingvistik leksikon — sinf	0,969
Psixolingvistik leksikon — hissiyot	0,948
Emoji hissiyoti	0,923

Alohida belgi qatlami detektorlarining held-out test natijalari (3-jadval) modelning empirik asosini to'ldiradi va forensik maqsadga mos xususiyatni — yuqori aniqlik, ehtiyotkor to'liqlik — ko'rsatadi.

3-jadval. Belgi qatlami detektorlarining held-out test ko'rsatkichlari

Qatlam / vazifa	P	R	F1
Imlo/grafika og'ishi	0,946	0,602	0,736
Frazeologik ibora aniqlash	—	—	0,928
Tuhmat (keng)	0,875	—	0,756
So'kinish niqobini ochish	0,936	—	0,75

Psixolingvistik qatlam detektorining mustaqil baholangan (oltin-600) natijalari 4-jadvalda keltirilgan. Neytral yangiliklar matnida soxta agressiya ulushi atigi 1,1% ni tashkil etdi — bu detektorning aniqlikka yo'naltirilganini va forensik maqsadlarga mosligini ko'rsatadi. Bilvosita agressiya va sarkazm leksik usulda deyarli aniqlanmaydi ($P = 0,158$), shu bois ular bo'yicha avtomatik hukm chiqarilmaydi.

4-jadval. Psixolingvistik qatlam detektorining held-out (oltin-600) natijalari

Maydon	κ	makro-F1	Qo'shimcha
Hissiyot	0,364	0,349	aniqlik 0,453
Tonallik	0,418	0,553	—
Agressiya	0,356	0,528	«bor/yo'q» P 0,842 / R 0,372

Muallifning jinsini ehtimoliy baholashda model held-out testda 0,567 aniqlik va 0,584 AUC ko'rsatdi; jins modelining o'z test AUC ko'rsatkichi 0,618 [0,589; 0,648], Cllr esa 0,962 ni tashkil etdi (1 dan kichik — model zaif, ammo informativ). Belgilar tanlovi (olmoshlar, xavotir leksikasi, kuchaytiruvchilar, murojaat shakllari, uzun so'zlar) xalqaro natijalarga mos keladi [11: 211; 12: 401], biroq ta'sir kattaligi kichik (Cohen $h \leq 0,37$) va bu janrda invektiv, tahdid hamda jargon markerlari jins bo'yicha sezilarli farq bermadi. Yosh guruhini aniqlashda yorliqlik tabiiy matnlar yetarli emas (qamrov ~4%), shu bois natija tasdiqlanmagan deb beriladi.

MUHOKAMA

Olingan natijalar kam resursli til va qisqa, nazoratli (sintetik) korpus sharoitida makro-F1 ko'rsatkichining 0,712 ga tengligi maqbul boshlang'ich natija ekanini tasdiqlaydi. DEV va TEST o'rtasidagi kichik farq modelning barqarorligini ko'rsatadi. Tahdid toifasidagi yuqori aniqlik va past to'liqlik o'zbek tilidagi tahdidlarning ko'pincha kinoya, shart yoki

yashirin ishora orqali ifodalanishi bilan izohlanadi; bu holat leksik qatlam ustiga kontekstni hisobga oluvchi klassifikator qatlamini qo'shishni taqozo etadi. Ablatsiya natijalari toksik nutqni aniqlashda yakka leksik markerlar yetarli emasligi haqidagi xalqaro xulosalarga mos keladi [10], belgi (subso'z) n-grammlaridan foydalanish esa morfologik boy tillar uchun afzal ekanligini tasdiqlaydi [9].

Muallif profilini baholashning kamtarona natijasi metodologik va etik jihatdan muhim saboq beradi: demografik va psixolingvistik markerlar ehtimoliy bo'lib, ular aniqlik darajasi bilan birga taqdim etilishi va hech qachon qat'iy hukm sifatida talqin qilinmasligi kerak [15; 11]. Tadqiqotning yana bir metodologik hissasi — oltin to'plam sifati (annotatorlararo κ) va detektor aniqligini qat'iy farqlashdir: yuqori κ yorliqlashdagi kelishuv darajasini, past κ esa annotatorlararo kelishuvning pastligini ifodalaydi va bu ko'rsatkich detektor aniqligi bilan aralashtirilmassligi kerak.

Modelni amaliyotga joriy etishning huquqiy-axloqiy jihatlari alohida e'tibor talab qiladi. Avtomatik tizimning natijasi qat'iy hukm emas, balki ehtimollik bilan ifodalangan dastlabki baho bo'lib, yakuniy xulosa malakali ekspert-tilshunos tomonidan tasdiqlanishi shart (inson nazorati tamoyili). Lingvistik profillash uchun foydalaniladigan matnlar shaxsiy ma'lumotlarni o'z ichiga olishi mumkin bo'lgani sababli, ularni yig'ish qonunchilik talablari doirasida, dalil butunligi (chain of custody) hujjatlashtirilgan holda amalga oshirilishi lozim; aynan shu sababli korpus to'liq sintetik shaklda yaratildi.

Natijalarni xalqaro va milliy tajriba bilan qiyoslaganda, markerga asoslangan izohlanuvchan yondashuv sud-ekspertiza konteksti uchun ikki jihatdan afzal ko'rinadi. Birinchidan, chiziqli SVM va marker-qoida modellarida har bir belgining qarorga qo'shgan hissasi shaffof tahlil qilinib, xulosada asoslab berilishi mumkin — bu "qora quti" neyron modellaridan farqli o'laroq, dalilning izohlanuvchanligi talabiga javob beradi. Ikkinchidan, avtomatlashtirilgan usulning sud dalili sifatidagi maqbulligi usulning sinovdan o'tganligi, xatolik darajasining ma'lumlighi va mutaxassislar hamjamiyatida e'tirof etilganligi mezonlariga bog'liq; muzlatilgan test va hujjatlashtirilgan xatolik darajasi (makro-F1 0,712) aynan shu mezonlarga xizmat qiladi. Shu bilan birga, chuqur o'rganish modellari kelajakda yashirin tahdid va sarkazmni aniqlashda qo'shimcha imkoniyatlar berishi mumkin, biroq ularning izohlanuvchanligi alohida tadqiqotni talab etadi.

Amaliy jihatdan bahsli matn ortidagi muallifning ehtimoliy kontekst toifasi va demografik profili sud-ekspertiza va tergov amaliyotida yo'naltiruvchi ahamiyatga ega: anonim tahdid yoki haqorat xati bo'yicha tekshiruv olib borilganda, lingvistik profil tekshiruv doirasini toraytirishga va yo'nalishni belgilashga ko'maklashadi. Ayni paytda, bu yondashuvning qo'llanilishi qat'iy metodologik chegaralarga bo'ysunadi: profil bahosi hech qachon yagona, hal qiluvchi dalil sifatida taqdim etilmaydi; u boshqa dalillar bilan tasdiqlanishi va malakali ekspert tomonidan talqin qilinishi shart. Shu ma'noda model tergovchilar uchun tezkor saralash vositasi, ekspertlar uchun esa lisoniy dalillarni baholashda yordamchi vositadir.

Tadqiqotning cheklovlari ochiq e'tirof etiladi: korpus sintetik bo'lgani uchun natijalar real dunyodagi taqsimotga emas, balki nazoratli sinovga tegishli va tabiiy korpusda qayta validatsiya zarur; barcha annotatsiyalar avtomatik bo'lib, inson-ekspert validatsiyasi kutilmoqda; marker qatlamlari yuqori aniqlik va nisbatan past to'liqlik bilan ishlaydi. Shunga qaramay, muzlatilgan test, yagona leksik reyestr, tekshirilgan iqtiboslar va izohlanuvchan qoidalarga tayangan jarayon natijalarning takrorlanuvchanligi hamda sud-ekspert xulosasida asoslanishi uchun zarur shaffoflikni ta'minlaydi; mustaqil tekshiruvchi barcha bosqichlarni qayta tasdiqladi.

Integratsiyalashgan belgi vektorining muhim afzalligi — modelning izohlanuvchanligini oshirishidir: tasniflagich qaror qabul qilganda, qaysi qatlam (uslubiy, demografik yoki individual markerlar) hal qiluvchi rol o'ynaganini ko'rsatish mumkin bo'ladi. Bu xususiyat sud-lingvistik ekspertizada alohida qimmatga ega, chunki ekspert xulosasi nafaqat yakuniy natijani, balki uning lingvistik asosini ham ochiq ko'rsatishi shart. Markerlar o'rtasidagi o'zaro moslik (masalan, yuqori emoji zichligi, norasmiy registr va muayyan harf n-grammlarining birgalikda kelishi) yakka belgidan ko'ra ishonchliroq, yaxlit dalil hosil qiladi.

XULOSA

Tadqiqotda o'zbek tilidagi bahsli matnlar uchun sintetik, etik jihatdan xavfsiz UCTS korpusi shakllantirildi, yetti qatlamli belgi fazosi va izohlanuvchan tasniflash modeli qurilib, muzlatilgan testda baholandi (kontekst makro-F1 0,712). Natijalar izohlanuvchan pragmatik qoidalarning ustunligini, muallif profillashining esa ehtimoliy va yordamchi tabiatini tasdiqlaydi. Tadqiqotning ilmiy yangiligi — o'zbek tilidagi bahsli matnlarga ixtisoslashgan korpus va baholash andozasi hamda oltin to'plam sifati bilan detektor aniqligini qat'iy farqlovchi metodologiyaning ishlab chiqilishidir. Istiqbolli yo'nalishlar — korpusni real matnlar bilan kengaytirish, yashirin tahdid uchun klassifikator qatlamini qo'shish, modelni inson-ekspert validatsiyasidan o'tkazish va to'liq o'rgatiladigan SVM tasniflagichi bilan qiyoslashdan iborat.

FOYDALANILGAN ADABIYOTLAR

1. Coulthard M., Johnson A. An Introduction to Forensic Linguistics: Language in Evidence. – London: Routledge, 2007. – 237 p.
2. Olsson J. Forensic Linguistics. – 2nd ed. – London: Continuum, 2008. – 294 p.
3. Stamatos E. A Survey of Modern Authorship Attribution Methods // Journal of the American Society for Information Science and Technology (JASIST). – 2009. – Vol. 60, № 3. – P. 538–556.
4. Koppel M., Schler J., Argamon S. Computational Methods in Authorship Attribution // Journal of the American Society for Information Science and Technology (JASIST). – 2009. – Vol. 60, № 1. – P. 9–26.
5. Biber D. Variation across Speech and Writing. – Cambridge: Cambridge University Press, 1988. – XIII, 299 p.
6. Lakoff R. Language and Woman's Place // Language in Society. – 1973. – Vol. 2, № 1. – P. 45–80.
7. Tannen D. You Just Don't Understand: Women and Men in Conversation. – New York: William Morrow, 1990.

8. Holmes J. Women, Men and Politeness. – London: Longman, 1995.
9. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching Word Vectors with Subword Information // Transactions of the Association for Computational Linguistics. – 2017. – Vol. 5. – P. 135–146.
10. Fortuna P., Nunes S. A Survey on Automatic Detection of Hate Speech in Text // ACM Computing Surveys. – 2018. – Vol. 51, № 4. – Article 85.
11. Newman M. L., Groom C. J., Handelman L. D., Pennebaker J. W. Gender Differences in Language Use // Discourse Processes. – 2008. – Vol. 45, № 3. – P. 211–236.
12. Koppel M., Argamon S., Shimon A. R. Automatically Categorizing Written Texts by Author Gender // Literary and Linguistic Computing. – 2002. – Vol. 17, № 4. – P. 401–412.
13. Nini A. A Theory of Linguistic Individuality for Authorship Analysis. – Cambridge: Cambridge University Press, 2023. – 96 p.
14. Grant T. Quantifying Evidence in Forensic Authorship Analysis // International Journal of Speech, Language and the Law (IJSLL). – 2007. – Vol. 14, № 1. – P. 1–25.
15. Searle J. R. Speech Acts: An Essay in the Philosophy of Language. – Cambridge: Cambridge University Press, 1969.
16. Safarov Sh. Pragmalingvistika. – Toshkent: O'zbekiston Milliy ensiklopediyasi, 2008. – 300 b.
17. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. – Cambridge, MA: MIT Press, 1999. – 680 p.
18. Cortes C., Vapnik V. Support-Vector Networks // Machine Learning. – 1995. – Vol. 20, № 3. – P. 273–297.
19. Platt J. Probabilistic Outputs for Support Vector Machines // Advances in Large Margin Classifiers. – Cambridge, MA: MIT Press, 1999. – P. 61–74.
20. Pedregosa F. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830.
21. Salton G., Buckley C. Term-weighting Approaches in Automatic Text Retrieval // Information Processing & Management. – 1988. – Vol. 24, № 5. – P. 513–523.
22. Spärck Jones K. A Statistical Interpretation of Term Specificity // Journal of Documentation. – 1972. – Vol. 28, № 1. – P. 11–21.
23. Cohen J. A Coefficient of Agreement for Nominal Scales // Educational and Psychological Measurement. – 1960. – Vol. 20, № 1. – P. 37–46.
24. Landis J. R., Koch G. G. The Measurement of Observer Agreement for Categorical Data // Biometrics. – 1977. – Vol. 33, № 1. – P. 159–174.
25. O'zbek tilining izohli lug'ati: 5 jildli. – Toshkent: O'zbekiston milliy ensiklopediyasi, 2006–2008.
26. O'zbek tilining izohli lug'ati: to'ldirilgan 6 jildli nashr / mas'ul muharrirlar A. Madvaliyev, D. Xudayberganova. – Toshkent: G'afur G'ulom nomidagi NMIU, 2020–2022.
27. Efron B., Tibshirani R. J. An Introduction to the Bootstrap. – New York: Chapman & Hall, 1993.
28. To'rayeva D. Mualliflikni aniqlashda sud-lingvistik profillash ekspertizasi metodikasi // Til va adabiyot.uz. – 2026. – № 11 (79). – B. 86–91.