

**O‘ZBEKISTON RESPUBLIKASI OLIY TA’LIM, FAN VA  
INNOVATSIYALAR VAZIRLIGI**

**MIRZO ULUG‘BEK NOMIDAGI O‘ZBEKISTON MILLIY  
UNIVERSITETINING JIZZAX FILIALI**



**KOMPYUTER ILMLARI VA MUHANDISLIK TEXNOLOGIYALARI**  
*mavzusidagi Xalqaro ilmiy-texnik anjuman materiallari to‘plami*  
**2026-yil 25-26-sentabr**

**(1-qism)**

**JIZZAX-2026**

Kompyuter ilmlari va muhandislik texnologiyalari. Xalqaro ilmiy-texnik anjuman materiallari to'plami – Jizzax: O'zMU Jizzax filiali, 2026-yil 25-26-sentabr. 507-bet.

Xalqaro miqyosidagi ilmiy-texnik anjuman materiallarida zamonaviy kompyuter ilmlari va muhandislik texnologiyalari sohasidagi innovatsion tadqiqotlar aks etgan.

Globalashuv sharoitida davlatimizni yanada barqaror va jadal sur'atlar bilan rivojlantirish bo'yicha amalga oshirilayotgan islohotlar samarasini yaxshilash sohasidagi ilmiy-tadqiqot ishlariga alohida e'tibor qaratilgan. Zero iqtisodiyotning, ijtimoiy sohalarni qamrab olgan modernizatsiya jarayonlari, hayotning barcha sohalarini liberallashtirishni talab qilmoqda.

Ushbu ilmiy ma'ruza tezislari to'plamida mamlakatimiz va xorijlik turli yo'nalishlarda faoliyat olib borayotgan mutaxassislar, olimlar, professor-o'qituvchilar, ilmiy tadqiqot institutlari va markazlarining ilmiy xodimlari, tadqiqotchilari, magistr va talabalarining ilmiy-tadqiqot ishlari natijalari mujassamlashgan.

Mas'ul muharrirlar: DSc.prof. Turakulov O.X., t.f.n., dots. Baboyev A.M.

Tahrir hay'ati a'zolari: p.f.d.(DSc), prof. Turakulov O.X., t.f.n., dots. Baboyev A.M., t.f.f.d.(PhD), prof. Abduraxmanov R.A., p.f.f.d.(PhD) Eshankulov B.S., p.f.n., dots. Aliqulov S.T., p.f.n., dots. Alimov N.N., p.f.f.d.(PhD), dots. Alibayev S.X., t.f.f.d.(PhD), dots. Abdumalikov A.A., p.f.f.d.(PhD) Hafizov E.A., f.f.f.d.(PhD) dots., Norbekova B.Sh. p.f.f.d.(PhD) dots. Sindorov L.K., t.f.f.d.(PhD), dots. Nasirov B.U., b.f.f.d. (PhD) O'ralov A.I., p.f.n., dots. Aliqulov S.T., t.f.f.d.(PhD) Kuvandikov J.T., i.f.n., dots. Tsoy M.P., Sharipova S.F., Jo'rayev M.M.

Mazkur to'plamga kiritilgan ma'ruza tezislarining mazmuni, undagi statistik ma'lumotlar va me'yoriy hujjatlarning to'g'riligi hamda tanqidiy fikr-mulohazalar, keltirilgan takliflarga mualliflarning o'zlari mas'uldirlar.

безопасность // Central Asian Journal of Education and Innovation. — 2024. — Т. 3, № 2. — С. 85–91. — DOI: 10.5281/zenodo.10644094.

9. Сайидов Ф.А., Хамдамов У.Р. Цифровая образовательная экосистема: анализ подходов к формированию единого информационного пространства для поддержки индивидуальных образовательных траекторий // Материалы междунар. конф. NamDPI. — 2026. — С. 252–259. — DOI: 10.5281/zenodo.19828824.

## **O‘ZBEK TILI IBORALARINI ANIQLASHDA LSA VA LDA GIBRID MODEL**

***Raxmatullaeva Nodira Baxodirovna***

*Toshkent shahridagi Belarus -O‘zbekiston qo‘shma tarmoqlararo amaliy texnik kvalifikatsiyalar instituti*

[\*nodirarahmatullayeva688@gmail.com\*](mailto:nodirarahmatullayeva688@gmail.com)

***Madetbayeva Nilufar Batirbayevna***

*Toshkent axborot texnologiyalari universiteti, Toshkent, O‘zbekiston*

[\*madetbaeva@gmail.com\*](mailto:madetbaeva@gmail.com)

**Annotatsiya:** Iborani gapda avtomatik topish mashinaviy tarjimaning dastlabki va hal qiluvchi bosqichidir. Ushbu ishda matn to‘plamidan avtomatik o‘rganiladigan ikki mavzu modeli, yashirin semantik tahlil (LSA) va yashirin Dirixle taqsimoti (LDA), bitta gibril belgilar to‘plamiga birlashtirilib, shu to‘plamning o‘zbek iboralari ontologiyasidan olingan belgilarga qo‘shimcha bera olishi tekshirildi. Sinov 860 ta belgilangan gap, 1,07 million leksemali matn to‘plami va 2 098 iborali ontologiya asosida uch xil tasniflagichda o‘tkazildi. Gibril mavzu belgilari o‘zi alohida olinganda F1 0,52 dan 0,57 gacha darajada qoldi, ontologiya belgilari ustiga qo‘shilganda esa uchala tasniflagichda ham natijani kamaytirdi va bu kamayish McNemar mezoni bo‘yicha ishonchli. Shu asosda hujjat darajasidagi mavzu modellari ibora aniqlash uchun yaroqsiz, ontologiya esa hal qiluvchi manba degan xulosa chiqariladi.

**Kalit so‘zlar:** ibora, ontologiya, yashirin semantik tahlil, yashirin Dirixle taqsimoti, gibril belgilar, tasniflash, o‘zbek tili, mashinaviy tarjima.

Sun‘iy intellekt texnologiyalarini 2030 yilga qadar rivojlantirish strategiyasi (Prezidentning 2024 yil 14 oktabrdagi PQ 358 sonli qarori) o‘zbek tili uchun til resurslari va aqlli dasturiy vositalar yaratishni davlat ahamiyatiga ega vazifa deb belgilaydi [1]. Bunday vositalar ichida mashinaviy tarjima alohida o‘rin tutadi va uning sifati ko‘p jihatdan iboralarga qanday munosabatda bo‘lishiga bog‘liq. Ibora tarkibidagi so‘zlar o‘z ma‘nosini yo‘qotib, birgalikda yangi ma‘no hosil qiladi [2]. Qo‘li ochiq birikmasi hech qanday qo‘l haqida emas, saxiy odam haqida, tarjimon esa buni gapdan avval ajratib olmasa, turk tiliga so‘zma so‘z o‘girib xato qiladi.

Avvalgi ishlarimizda iboralarni ajratish uchun o‘zbek iboralari ontologiyasi tuzilgan va undan olinadigan belgilar tasniflagichlarda yaxshi natija bergan edi [3].

Ammo ontologiya yopiq ro'yxat, u yangi yoki o'zgargan shaklni bilmaydi. Shuning uchun matn to'plamining o'zidan, hech qanday qo'l mehnatisiz o'rganiladigan belgilar ontologiya kamchiligini qoplay oladimi degan savol o'rta qo'yiladi. Bu tezisda shu savolga mavzu modellari nuqtai nazaridan javob izlanadi. Ikki usul tanlandi, ular bir xil narsani ikki xil matematik tilda qiladi, LSA hujjat va so'z matritsasini chiziqli algebra yo'li bilan siqadi [4], LDA esa hujjatni bir necha mavzuning ehtimolliq aralashmasi deb tushuntiradi [5]. Ikkalasini bitta gibrid vektorga jamlab, ontologiya belgilari ustiga qo'shamiz va farqni o'lchaymiz.

Birinchi usulda matn to'plami  $W$  matritsa ko'rinishiga keltiriladi, uning satrlari hujjatlar, ustunlari so'zlar, elementlari TF-IDF og'irliklari. Matritsa singulyar yoyilma orqali  $k$  ta asosiy o'qqa qisqartiriladi va har bir yangi gap shu o'qlar bo'yicha koordinatalarga ega bo'ladi.

$$W \approx U_k \Sigma_k V_k^T, \quad \hat{d} = \Sigma_k^{-1} U_k^T d \quad (1)$$

Bu yerda:

$W$  – hujjatlar va so'zlar bo'yicha TF-IDF og'irliklar matritsasi

$U_k, \Sigma_k, V_k$  – singulyar yoyilmaning dastlabki  $k$  o'qi

$d$  – sinlayotgan gapning og'irliklar vektori

$\hat{d}$  – shu gapning  $k$  ta o'q bo'yicha koordinatalari

Ikkinchi usulda har bir gap uchun mavzular ulushi  $\hat{\theta}_d$  o'rgatilgan LDA modelidan olinadi. O'rgatishda har bir so'zga qaysi mavzu tegishli ekani ma'lum emas, shu sababli u Gibbs tanlab olish yo'li bilan qayta qayta baholanadi [6].

$$p(z_i = t \mid \mathbf{z}_{-i}, \mathbf{w}) \propto (n_{d,t}^{-i} + \alpha) \frac{n_{t,w_i}^{-i} + \beta}{n_t^{-i} + V\beta} \quad (2)$$

Bu yerda:

$z_i - i$  o'rindagi so'zga berilayotgan mavzu

$n_{d,t}^{-i}$  – shu hujjatda  $t$  mavzuga tegishli deb topilgan so'zlar soni,  $i$  so'z hisobga olinmagan holda

$n_{t,w_i}^{-i}$  –  $w_i$  so'zning butun to'plamda  $t$  mavzuga tegishli deb topilgan soni

$n_t^{-i}$  –  $t$  mavzuga tegishli barcha so'zlar soni

$\alpha, \beta$  – Dirixle taqsimotining oldindan berilgan parametrlari,  $V$  – lug'atdagi so'zlar soni

Gibrid model ikki usulning chiqishini leksik va ontologik belgilar bilan bitta qatorga tizadi va tasniflagichga uzatadi.

$$x_d = [x_d^{lex}, x_d^{onto}, \hat{d}^{LSA}, \hat{\theta}_d], \quad \hat{y}_d = \text{sign}(g(x_d)) \quad (3)$$

Bu yerda:

$x_d^{lex}$  – gapdagi leksemalar va ularning juftliklari bo'yicha TF-IDF og'irliklari

$x_d^{onto}$  – ontologiyadan olingan sakkiz belgi, jumladan tayanch va tarkibiy leksemalarning uchrash soni, ularning gap uzunligiga nisbati, eng ko'p qamrab olingan ibora ulushi, to'liq mos kelgan iboralar soni

$\hat{d}^{LSA}, \hat{\theta}_d$  – 50 ta LSA koordinatasi va 20 ta LDA mavzu ulushi

$g$  – tasniflagich funksiyasi,  $\hat{y}_d$  – gapda ibora bor yoki yo'q degan qaror

Ma'lumotlar quyidagicha. Belgilangan to'plam 860 gapni o'z ichiga oladi, ularning 399 tasi iborali, 461 tasi iborasiz, gaplar sakkiz badiiy asardan tanlab

olingan. Mavzu modellari 154 502 gapli leksemashtirilgan matn to'plamida o'rgatildi, bunda o'nta ketma-ket gap bitta hujjat deb qaraldi, jami 15 347 hujjat va 10 809 leksemali lug'at hosil bo'ldi, belgilangan 860 gap bu jarayondan chetlatildi. Ontologiyadan 2 098 iboraning tarkibi, 466 tayanch va 1 795 tarkibiy leksema olindi [3]. Mavzular soni izchillik o'lchovi bo'yicha 20 deb tanlandi [7], LSA uchun 50 o'q olindi. Tasniflagich sifatida chiziqli SVM, gradient boosting va logistik regressiya ishlatildi. Har bir variant besh bo'lakli kesishma tekshiruvda beshta turli boshlang'ich qiymat bilan sinaldi, ya'ni har biri uchun 25 ta o'lchov olindi, farqlar McNemar mezoni bilan tekshirildi.

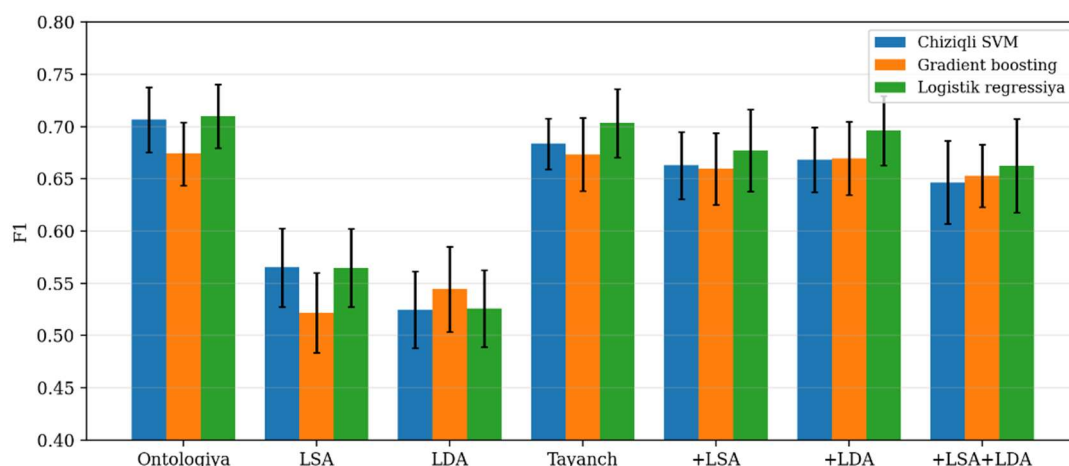
Ikki dastlabki kuzatuv natijani oldindan tushuntiradi. LSA da 50 ta o'q butun matritsa o'zgaruvchanligining faqat 7,8 foizini, 100 ta o'q 12,9 foizini qamrab oladi, ya'ni badiiy matnda ma'no oz sonli kuchli o'qqa yig'ilmaydi. LDA mavzulari esa iboraga emas, asar va qahramonlarga bog'lanib chiqdi, masalan bir mavzuning eng ehtimolli so'zlari yo'lchi, gulnor, boy, yormat, xo'jayin, boshqasini oy, qiz, xotin, ra'no, opa. Bunday mavzu gapda ibora bor yoki yo'qligi haqida deyarli hech narsa aytmaydi.

Asosiy natijalar 1-rasmda jamlangan. Faqat mavzu belgilari bilan tasniflagichlar F1 bo'yicha 0,52 dan 0,57 gacha darajaga chiqdi, to'plamda ikki sinf deyarli teng bo'lgani uchun bu tasodifiy bashoratdan ozgina yuqori. Faqat ontologiya belgilari bilan esa F1 0,67 dan 0,71 gacha. Leksik va ontologik belgilardan iborat tayanch model ustiga gibrid mavzu vektori qo'shilganda F1 chiziqli SVM da 0,684 dan 0,647 ga, gradient boostingda 0,673 dan 0,653 ga, logistik regressiyada 0,703 dan 0,663 ga tushdi, uchala holatda ham McNemar p qiymati 0,001 dan kichik. LSA va LDA alohida qo'shilganda ham manzara o'zgarmadi, faqat gradient boostingda LDA qo'shilishi neytral chiqdi ( $p = 0,389$ ).

| Belgilar to'plami            | SVM F1            | SVM p   | GB F1             | GB p    | LR F1             | LR p    |
|------------------------------|-------------------|---------|-------------------|---------|-------------------|---------|
| Faqat ontologiya             | 0,706 $\pm$ 0,031 | < 0,001 | 0,674 $\pm$ 0,030 | 0,342   | 0,710 $\pm$ 0,030 | 0,129   |
| Faqat LSA                    | 0,565 $\pm$ 0,038 | < 0,001 | 0,522 $\pm$ 0,038 | < 0,001 | 0,565 $\pm$ 0,037 | < 0,001 |
| Faqat LDA                    | 0,525 $\pm$ 0,037 | < 0,001 | 0,544 $\pm$ 0,041 | < 0,001 | 0,526 $\pm$ 0,037 | < 0,001 |
| Tayanch (leksik+onto)        | 0,684 $\pm$ 0,024 |         | 0,673 $\pm$ 0,035 |         | 0,703 $\pm$ 0,033 |         |
| Tayanch + LSA                | 0,663 $\pm$ 0,032 | < 0,001 | 0,660 $\pm$ 0,035 | 0,025   | 0,677 $\pm$ 0,039 | < 0,001 |
| Tayanch + LDA                | 0,668 $\pm$ 0,031 | 0,002   | 0,670 $\pm$ 0,035 | 0,389   | 0,696 $\pm$ 0,033 | 0,023   |
| Tayanch + LSA + LDA (gibrid) | 0,647 $\pm$ 0,040 | < 0,001 | 0,653 $\pm$ 0,030 | < 0,001 | 0,663 $\pm$ 0,045 | < 0,001 |

*1-rasm. Belgilar to'plamlari bo'yicha F1 natijalari (25 o'lchov o'rtachasi va standart chetlanishi, p tayanch modelga nisbatan)*

2-rasm uchala tasniflagichni yonma yon ko'rsatadi. Hech bir ustunda mavzu belgisi qo'shilgan variant tayanchdan baland emas, gibrid variant esa har uch tasniflagichda eng pastda.



*2-rasm. Uchala tasniflagichda belgilar to‘plamlarining F1 bo‘yicha taqqoslanishi*

Natijani uch omil tushuntiradi. Ibora gap ichida bir necha so‘zlik tor joyda yashaydi va so‘zlar tartibiga bog‘liq, mavzu modeli esa o‘nlab gapdan tuzilgan hujjatga qaraydi va so‘zlarni tartibsiz to‘plam deb hisoblaydi. Mavzu vektori har bir gap uchun o‘rtacha yettita leksema asosida hisoblanadi, bu juda oz va beqaror. Nihoyat, tasniflagichga qo‘shilgan 70 ta yangi koordinata foydali axborot emas, asar va qahramon haqidagi shovqin olib keladi, chiziqli modellar undan ayniqsa zarar ko‘radi, daraxtli model esa uni qisman chetlab o‘ta oladi.

Demak LSA va LDA dan tuzilgan gibrid mavzu belgilari o‘zbek tili iboralarini aniqlashda ontologiya belgilarining o‘rnini bosa olmaydi va ularga qo‘shimcha ham bermaydi. Bu ontologiyani ibora aniqlashdagi markaziy o‘rnini eksperimental tarzda tasdiqlaydi. Keyingi bosqichda so‘z darajasida ishlaydigan vektor modellarini ontologiya bilan birlashtirish sinaladi, chunki ular hujjat emas, so‘z atrofidagi tor kontekstga qaraydi.

#### **Foydalanilgan adabiyotlar ro‘yxati:**

1. O‘zbekiston Respublikasi Prezidentining 2024 yil 14 oktabrdagi PQ 358 sonli Sun’iy intellekt texnologiyalarini 2030 yilga qadar rivojlantirish strategiyasini tasdiqlash to‘g‘risidagi qarori. lex.uz.
2. Baldwin T., Kim S.N. Multiword expressions. Handbook of Natural Language Processing, 2nd ed. Boca Raton, CRC Press, 2010, pp. 267–292.
3. Nazirova E., Abdurakhmonova N., Madetbayeva N. O‘zbek tili iborolari ontologiyasi va o‘xshashlik o‘lchovlari. UBMK 2026.
4. Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 1990, pp. 391–407.
5. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 2003, pp. 993–1022.
6. Griffiths T.L., Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences, 101 (suppl. 1), 2004, pp. 5228–5235.
7. Mimno D., Wallach H., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models. Proceedings of EMNLP 2011, Edinburgh, 2011, pp. 262–272.