DAVLAT TASHKILOTLARI HUJJATLARINI SUN’IY INTELLEKT ASOSIDA QAYTA ISHLASH VA SEMANTIK QIDIRUV TIZIMLARINING NAZARIY ASOSLARI

Nasrullayev Nurbek Baxtiyorovich
Muhammad al Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti
Kiberxavfsizlik fakulteti dekani
n.nasrullayev@tuit.uz

Saydaxmedov Ibodilla Xikmatullayevich
Tashkent International University of Education
Magistrant
ibodillakhodja@gmail.com

Hamzayev Jamshid Fayzidin o‘g‘li
Muhammad al Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti
Kompyuter tizimlari kafedrasi dotsenti
hamzayevjf@gmail.com
Tel: (+998997873764)

Annotatsiya. *Ushbu maqolada davlat tashkilotlari faoliyatida shakllanadigan hujjatlarni sun’iy intellekt texnologiyalari asosida avtomatik qayta ishlash hamda semantik qidiruv tizimlarini yaratishning nazariy-uslubiy asoslari tadqiq etilgan. Tadqiqotning dolzarbligi shundan iboratki, davlat tashkilotlarida yig‘ilgan hujjatlar hajmining tobora ortib borishi sharoitida an’anaviy, kalit so‘zga asoslangan qidiruv usullari foydalanuvchi so‘roviga mazmunan mos, ammo leksik jihatdan farqli hujjatlarni aniqlay olmaydi. Maqolaning maqsadi bu elektron hujjat aylanish tizimlarining zamonaviy holatini, tabiiy tilni qayta ishlash (NLP) usullarini hamda embedding modellari va vektorli ma’lumotlar bazalarining ishlash tamoyillarini tizimli tahlil qilish va davlat sektori uchun mos kontseptual arxitekturani asoslashdan iborat. Tadqiqot metodologiyasi tizimli adabiyotlar sharhi, qiyosiy tahlil va kontseptual modellashtirish usullarini o‘z ichiga oladi. Tahlil uchun 2013-2024-yillarda chop etilgan ilmiy manbalar, xalqaro standartlar va milliy me’yoriy-huquqiy hujjatlar tanlab olindi. Tahlil natijasida mavjud elektron hujjat aylanish tizimlarida sun’iy intellekt integratsiyasi cheklanganligi, transformer arxitekturasiga asoslangan zamonaviy NLP modellari matn mazmunini yuqori aniqlikda ifodalash imkonini berishi hamda embedding va vektorli ma’lumotlar bazalari semantik qidiruvni amalga oshirishning asosiy texnik vositasi ekanligi aniqlandi. Tahlil asosida besh qatlamli (manba, OCR, NLP, embedding yoki vektorli saqlash, qidiruv interfeysi) kontseptual arxitektura taklif etilgan. Natijalar shuni ko‘rsatadiki, davlat tashkilotlari uchun milliy tilga moslashtirilgan semantik qidiruv tizimini yaratish amaliy jihatdan asoslangan va dolzarb yo‘nalish hisoblanadi.*

Kalit so‘zlar: *sun’iy intellekt, davlat tashkiloti, elektron hujjat aylanishi, tabiiy tilni qayta ishlash, semantik qidiruv, embedding, vektorli ma’lumotlar bazasi, matnni klassifikatsiyalash, raqamli transformatsiya, axborot izlash tizimlari.*

KIRISH

Zamonaviy davlat boshqaruvi tizimida axborotning hajmi va murakkabligi jadal sur’atlarda ortib bormoqda. Davlat tashkilotlari kundalik faoliyati davomida qonun

hujjatlari, buyruqlar, xatlar, arizalar, hisobotlar va boshqa turdagi ko‘plab matnli hujjatlarni yaratadi, qabul qiladi va arxivlaydi. O‘zbekiston Respublikasida “Raqamli O‘zbekiston - 2030” strategiyasi doirasida davlat xizmatlarini raqamlashtirish, elektron hukumat tizimlarini rivojlantirish va davlat boshqaruvida sun’iy intellekt texnologiyalarini joriy etish ustuvor vazifa sifatida belgilangan [1]. Shunga qaramay, davlat tashkilotlarida to‘plangan yuz minglab hujjatlar orasidan kerakli ma’lumotni tezkor va aniq topish muammosi hamon dolzarbligicha qolmoqda.

Mavjud elektron hujjat aylanish tizimlari (EHAT), asosan, hujjatlarni ro‘yxatga olish, ijro nazorati va idoralararo marshrutlashni avtomatlashtirishga yo‘naltirilgan bo‘lib, ularning mazmunini chuqur tahlil qilish imkoniyati juda cheklangan [2]. Tadqiqotchilar ta’kidlashicha, bunday tizimlarda qidiruv, odatda, aniq kalit so‘zlarga asoslangan mexanizm orqali amalga oshiriladi, bu esa foydalanuvchi so‘rovi bilan mazmunan yaqin, ammo leksik jihatdan farqli hujjatlarni topib bera olmasligiga olib keladi [3]. Natijada, davlat xizmatchilari zarur hujjatni qidirishga ortiqcha vaqt sarflaydi, bu esa qaror qabul qilish jarayonining samaradorligiga salbiy ta’sir ko‘rsatadi.

So‘nggi yillarda sun’iy intellekt, xususan, tabiiy tilni qayta ishlash (Natural Language Processing - NLP) va chuqur o‘rganish texnologiyalarining rivojlanishi matnli ma’lumotlarni semantik darajada tahlil qilish imkonini yaratdi. Transformer arxitekturasi va unga asoslangan BERT kabi oldindan o‘qitilgan til modellarining paydo bo‘lishi matnlarni vektorli fazoga (embedding) aylantirish va ular orasidagi semantik o‘xshashlikni matematik jihatdan aniq hisoblash imkonini berdi [4, 5]. Bu o‘z navbatida, mazmunga yo‘naltirilgan, ya’ni semantik qidiruv tizimlarining shakllanishiga asos bo‘ldi.

Ushbu tadqiqotning maqsadi bu davlat tashkilotlari hujjatlarini sun’iy intellekt asosida qayta ishlash va semantik qidiruv tizimlarini yaratishning nazariy asoslarini tizimli tahlil qilish hamda ularning kontseptual arxitekturasini asoslashdan iborat. Maqsadga erishish uchun quyidagi vazifalar belgilandi:

1) Davlat tashkilotlarida elektron hujjat aylanish tizimlarining xususiyatlarini tahlil qilish;

2) Matnli ma’lumotlarni sun’iy intellekt asosida qayta ishlash usullari va algoritmlarini ko‘rib chiqish;

3) Semantik qidiruv tizimlari, embedding modellari va vektorli ma’lumotlar bazalarining zamonaviy usullarini tahlil qilish.

Tadqiqot obyekti sifatida davlat tashkilotlarining elektron hujjat aylanish jarayonlari, predmeti sifatida esa ushbu jarayonlarni sun’iy intellekt vositalari yordamida avtomatlashtirish va semantik qidiruvni tashkil etish usullari belgilandi. Tadqiqotning ilmiy yangiligi shundan iboratki, unda birinchi marta davlat sektori hujjatlari uchun OCR, NLP, embedding va vektorli qidiruvni yagona zanjirga birlashtiruvchi besh qatlamli kontseptual arxitektura taklif etilmoqda.

ADABIYOTLAR SHARHI

Tadqiqot mavzusi bo‘yicha mavjud ilmiy adabiyotlarni uchta asosiy yo‘nalish bo‘yicha tizimlashtirish maqsadga muvofiq topildi. Ya’ni elektron hujjat aylanish

tizimlari, matnli ma’lumotlarni sun’iy intellekt asosida qayta ishlash usullari hamda semantik qidiruv va vektorli texnologiyalar.

Xalqaro amaliyotda hujjatlarni boshqarish standartlari, xususan ISO 15489-1:2016 standarti, tashkilotlarda hujjatlarni yaratish, saqlash va yo‘q qilish jarayonlariga qo‘yiladigan asosiy talablarni belgilaydi [6], biroq ushbu standart hujjat mazmunini avtomatik tahlil qilish yoki semantik qidiruvni tashkil etish bo‘yicha aniq mexanizmlarni nazarda tutmaydi. O‘zbekiston Respublikasida davlat tashkilotlarida elektron hujjat aylanishini joriy etish bo‘yicha me’yoriy-huquqiy baza shakllangan bo‘lsada [2], milliy tadqiqotchilar ta’kidlashicha, amaldagi tizimlar asosan hujjat maqomini kuzatish va ijro intizomini ta’minlashga xizmat qiladi, mazmuniy tahlil funksiyalari esa deyarli rivojlantirilmagan [3]. Xalqaro tijorat yechimlari (SharePoint, OpenText, Alfresco) tahlili shuni ko‘rsatadiki, sun’iy intellekt imkoniyatlari ularda, odatda, qo‘shimcha modul yoki tashqi integratsiya sifatida taqdim etiladi va milliy tilga moslashtirilmagan holda yetkazib beriladi, bu esa davlat sektori uchun mahalliylashtirilgan yechimlarga bo‘lgan ehtiyojni kuchaytiradi.

Shuningdek, xorijiy tadqiqotchilar tomonidan o‘tkazilgan qiyosiy tahlillar shuni ko‘rsatadiki, elektron hujjat boshqaruvi tizimlarini joriy etishning muvaffaqiyati ko‘p jihatdan tashkilotning axborot infratuzilmasi tayyorligi va xodimlarning raqamli savodxonlik darajasiga bog‘liq, bu esa sof texnologik yechimlar bilan bir qatorda tashkiliy omillarni ham hisobga olish zarurligini ko‘rsatadi.

Matnni sonli ko‘rinishga o‘tkazish usullarining rivojlanish tarixi statistik yondashuvlardan (TF-IDF, Bag-of-Words) neyron tarmoqqa asoslangan usullarga qarab evolyutsiya qilgan. Mikolov va boshqalar tomonidan taklif etilgan Word2Vec modeli so‘zlarni past o‘lchamli vektor fazoga joylashtirish orqali ularning semantik yaqinligini ifodalash imkonini berdi [7], biroq bunday modellar kontekstni, xususan bir so‘zning turli ma’nolarini (polisemiya) hisobga ololmaydi. Ushbu cheklovni bartaraf etishda transformer arxitekturasining taklif etilishi burilish nuqtasi bo‘ldi [4]. Ya’ni o‘z-o‘ziga e’tibor (self-attention) mexanizmi matndagi uzoq masofali bog‘liqliklarni samarali aniqlash imkonini berdi. Ushbu arxitektura asosida yaratilgan BERT [5] va keyinchalik GPT oilasiga mansub yirik til modellari [8] matnni tushunish va generatsiya qilish vazifalarida sezilarli yutuqlarga erishdi. Shu bilan birga, tadqiqotchilar ta’kidlaydiki, bunday modellarni resurslari nisbatan kam bo‘lgan tillarga, jumladan o‘zbek tiliga moslashtirish alohida lingvistik va texnik yondashuvni talab qiladi [9]. Klassik pattern recognition va mashinali o‘rganish nazariyasi nuqtai nazaridan, matnni klassifikatsiyalash vazifasi umumiy nazorat ostidagi o‘qitish muammosining xususiy holati sifatida ko‘rib chiqiladi [10], bu esa hujjat turlarini avtomatik aniqlashda klassik va neyron usullarni birgalikda qo‘llash imkoniyatini asoslaydi.

Keltirilgan tadqiqotlarni umumlashtirish shuni ko‘rsatadiki, klassik statistik usullar hisoblash resursi jihatidan tejamkor bo‘lsada, semantik bog‘liqlikni umuman hisobga olmaydi, neyron so‘z vektorlari qisman semantik yaqinlikni ifodalaydi ammo kontekstga sezgir emas, transformer asosidagi modellar esa yuqori aniqlikka erishadi, biroq katta hisoblash quvvati va katta hajmdagi o‘qitish ma’lumotlarini talab qiladi. Ushbu tendensiya aniqlik va resurs sarfi o‘rtasidagi murosani ko‘rsatadi, bu davlat

tashkilotlari kabi resurslari cheklangan muassasalar uchun model tanlashda hal qiluvchi omil hisoblanadi.

Jumla darajasidagi embeddinglarni samarali hisoblash muammosi Reimers va Gurevych tomonidan taklif etilgan Sentence-BERT (SBERT) modelida siyam (Siamese) tarmoq arxitekturasi yordamida hal qilingan, bu esa yuzlab minglab jumlalar orasidan kosinusli o‘xshashlik bo‘yicha tezkor taqqoslash imkonini berdi [11]. Ko‘p tilli muhitda, jumladan resurslari nisbatan kam tillarda ishlash uchun Wang va boshqalar tomonidan taklif etilgan multilingual-E5 kabi modellar samarali yechim sifatida ko‘rsatilgan [12]. Millionlab hujjatlar orasidan tezkor qidiruvni ta’minlash muammosi Johnson va boshqalar tomonidan ishlab chiqilgan FAISS kutubxonasida GPU asosidagi taqribiy qidiruv algoritmlari yordamida hal etilgan [13], keyinchalik Malkov va Yashunin tomonidan taklif etilgan HNSW (Hierarchical Navigable Small World) grafik strukturasi qidiruv tezligi va aniqligi o‘rtasidagi muvozanatni sezilarli darajada yaxshiladi [14]. Axborot izlashning klassik nazariy asoslari Manning va boshqalar [15], shuningdek Salton va McGill tomonidan [16] shakllantirilgan bo‘lib, ular zamonaviy vektorli qidiruv tizimlarining matematik negizini tashkil etadi. So‘nggi yillarda Lewis va boshqalar tomonidan taklif etilgan Retrieval-Augmented Generation (RAG) yondashuvi semantik qidiruvni generativ til modellari bilan birlashtirib, foydalanuvchiga nafaqat hujjatni topish, balki uning mazmuni asosida tabiiy tilda javob shakllantirish imkonini berdi [17]. FAISS kutubxonasining so‘nggi texnik tavsifi shuni ko‘rsatadiki, bunday vositalar endilikda sanoat miqyosida, millionlab va hatto milliardlab vektorlar bilan ishlashga moslashtirilgan [18].

Keltirilgan adabiyotlar tahlili shuni ko‘rsatadiki, elektron hujjat aylanishi, matnli ma’lumotlarni qayta ishlash hamda semantik qidiruv yo‘nalishlarining har biri alohida-alohida yetarlicha chuqur o‘rganilgan. Shu bilan birga, ushbu uch yo‘nalishni davlat sektori ehtiyojlariga moslashtirilgan holda, milliy tilni hisobga olgan holda yagona tizimga birlashtiruvchi kompleks tadqiqotlar amalda deyarli uchramaydi. Aynan shu bo‘shliq ushbu tadqiqotning dolzarbligini belgilaydi va uni to‘ldirish maqsadida quyida taklif etilayotgan kontseptual arxitektura ishlab chiqilgan.

METODOLOGIYA

Tadqiqot nazariy-analitik xarakterga ega bo‘lib, tizimli adabiyotlar sharhi elementlaridan foydalangan holda olib borildi. Tadqiqot dizayni uch bosqichdan iborat qilib tuzildi: manbalarni izlash va tanlash, qiyosiy tahlil hamda kontseptual modellashtirish.

Manbalarni izlash jarayonida Google Scholar, Scopus, IEEE Xplore va ResearchGate ilmiy bazalari, shuningdek O‘zbekiston Respublikasi qonun hujjatlari ma’lumotlar bazasi (lex.uz) dan foydalanildi. Qidiruv “electronic document management system”, “semantic search”, “text embedding”, “natural language processing government”, “vector database”, “elektron hujjat aylanishi”, “sun’iy intellekt davlat boshqaruvi” kabi kalit so‘z birikmalari bo‘yicha amalga oshirildi. Vaqt oralig‘i sifatida 2013-2024-yillar belgilandi, bunda asosiy e’tibor 2018-yildan keyin chop etilgan manbalarga qaratildi, chunki aynan shu davrda transformer arxitekturasiga asoslangan modellar keng tarqaldi.

Manbalarni tanlashda quyidagi kiritish mezonlaridan foydalanildi: retsenziyalangan ilmiy jurnal yoki xalqaro konferensiya materiali, rasmiy standart yoki me’yoriy-huquqiy hujjat bo‘lishi, mavzuga bevosita aloqadorlik, to‘liq matnning ochiq yoki institutsional kirish orqali mavjudligi. Chiqarish mezonlari sifatida quyidagilar belgilandi: faqat annotatsiya darajasida mavjud bo‘lgan ishlar, mazmunan takrorlanuvchi (dublikat) nashrlar, shuningdek metodologik jihatdan yetarlicha asoslanmagan ommabop manbalar.

Dastlabki qidiruv natijasida taxminan 60 dan ortiq potentsial manba aniqlandi. Sarlavha va annotatsiya bo‘yicha skrining bosqichidan so‘ng ularning soni 35 taga qisqartirildi. To‘liq matnni tahlil qilish natijasida yakuniy tahlilga jami 18 ta manba kiritildi, ular orasida ilmiy maqolalar, monografiyalar, xalqaro standart hamda milliy me’yoriy-huquqiy hujjatlar mavjud.

Tanlangan manbalar asosida qiyosiy tahlil o‘tkazildi. Elektron hujjat aylanish tizimlarini solishtirish uchun qamrov darajasi, sun’iy intellekt integratsiyasi va asosiy cheklovlar mezonlaridan foydalanildi (1-jadval). Matnli ma’lumotlarni qayta ishlash usullarini solishtirishda ishlash tamoyili, afzalliklari va kamchiliklari mezonlari qo‘llanildi (2-jadval). Embedding modellari va vektorli ma’lumotlar bazalarini tahlil qilishda esa model turi, vektor o‘lchami va qo‘llanish sohasi mezonlari asos qilib olindi (3-jadval). Mezonlar tanlovi ilgari o‘tkazilgan analogik qiyosiy tadqiqotlar tajribasiga tayangan holda amalga oshirildi.

Qiyosiy tahlil natijalari asosida kontseptual modellashtirish bosqichi amalga oshirildi. Ushbu bosqichda avval hujjatni qayta ishlashning mantiqiy bosqichlari (manba, matn ekstraksiyasi, tilni qayta ishlash, vektorlashtirish, qidiruv) alohida-alohida ajratildi, so‘ngra ular orasidagi funksional bog‘liqliklar aniqlandi va besh qatlamli arxitektura ko‘rinishida rasmiylashtirildi (3-rasm). Modelni vizual tarzda ifodalash uchun blok-sxema usulidan foydalanildi, bu esa arxitekturaning har bir komponentini alohida va bir-biriga bog‘liq holda ko‘rsatish imkonini berdi.

Tadqiqot natijalarining ishonchliligini ta’minlash maqsadida manbalarning turli xil tipologiyasi (ilmiy nashrlar, xalqaro standartlar, milliy me’yoriy-huquqiy hujjatlar) bo‘yicha triangulyatsiya qo‘llanildi, ya’ni bir xil xulosa kamida ikki mustaqil manba turi bilan tasdiqlanishiga harakat qilindi. Shuningdek, jadvallarda keltirilgan qiyosiy ma’lumotlar bir necha marta manba matnlari bilan qayta solishtirildi.

Tadqiqot metodologiyasining bir qator cheklovlari mavjud. Birinchidan, tahlilga faqat ochiq yoki institutsional kirish imkoni bo‘lgan manbalar kiritildi, bu ba’zi tijorat tizimlari bo‘yicha to‘liq texnik hujjatlarning tahlildan chetda qolishiga sabab bo‘lgan bo‘lishi mumkin. Ikkinchidan, ushbu maqola doirasida taklif etilgan kontseptual arxitekturaning amaliy dasturiy sinovi (empirik baholash) o‘tkazilmadi bu keyingi tadqiqot bosqichida, dissertatsiyaning amaliy bobida amalga oshirilishi rejalashtirilgan.

TAHLIL VA NATIJALAR

O‘tkazilgan qiyosiy tahlil natijalari uch yo‘nalish ya’ni elektron hujjat aylanish tizimlari, matnli ma’lumotlarni qayta ishlash usullari hamda semantik qidiruv texnologiyalari bo‘yicha quyida keltirilgan.

1-jadvalda keltirilgan qiyosiy tahlil natijalariga ko‘ra, davlat sektorida foydalaniladigan milliy EHAT tizimi asosan idoralararo hujjat aylanishini qamrab oladi va marshrutlashni avtomatik amalga oshiradi, biroq sun’iy intellekt imkoniyati qisman, faqat avtomatik marshrutlash darajasida mavjud. Tijorat yechimlaridan SharePoint kuchli SI integratsiyasiga (Copilot, Graph API) ega, biroq litsenziya narxi yuqori va davlat buluti infratuzilmasini talab qiladi. OpenText Content Suite qoidaga asoslangan o‘rtacha darajadagi integratsiyani ta’minlaydi, Alfresco esa ochiq kodli yechim sifatida qo‘shimcha modul orqali SI imkoniyatlarini taqdim etadi.

1-jadval.

Elektron hujjat aylanish tizimlarining sun’iy intellekt integratsiyasi bo‘yicha qiyosiy tahlili

Tizim nomi	Qamrov darajasi	SI integratsiyasi	Asosiy cheklovlar
“E-Hujjat” DEHAT (O‘zbekiston)	Vazirlik va idoralararo hujjat aylanishi	Qisman (avtomatik marshrutlash)	Semantik qidiruv va mazmuniy tahlil imkoniyati cheklangan
SharePoint / Microsoft 365	Korporativ hujjat boshqaruvi	Kuchli (Copilot, Graph API)	Litsenziya narxi yuqori, davlat buluti talab qilinadi
OpenText Content Suite	Yirik tashkilotlar uchun ECM	O‘rtacha (qoidaga asoslangan)	Moslashtirish murakkab, mahalliylashtirish talab etadi
Alfresco (ochiq kodli)	O‘rta va yirik tashkilotlar	Qo‘shimcha modul orqali (plagin)	Sozlash uchun malakali IT-xodim zarur
Taklif etilayotgan model	Davlat tashkiloti hujjatlari, ko‘p tilli	To‘liq (NLP + embedding + vektorli qidiruv)	Ilmiy-tadqiqot bosqichida, sinov muhitida sinovdan o‘tkazilishi lozim

2-jadvalda matnli ma’lumotlarni qayta ishlashda qo‘llaniladigan olti asosiy usulning qiyosiy tahlili keltirilgan. Tahlil natijalariga ko‘ra, TF-IDF va Word2Vec/GloVe kabi statistik va erta neyron usullar hisoblash jihatidan tejamkor, biroq semantik bog‘liqlikni cheklangan darajada ifodalaydi. RNN/LSTM arxitekturalari ketma-ketlikni qisman kontekstual saqlaydi, ammo uzoq ketma-ketliklarda gradientning yo‘qolishi muammosiga duch keladi.

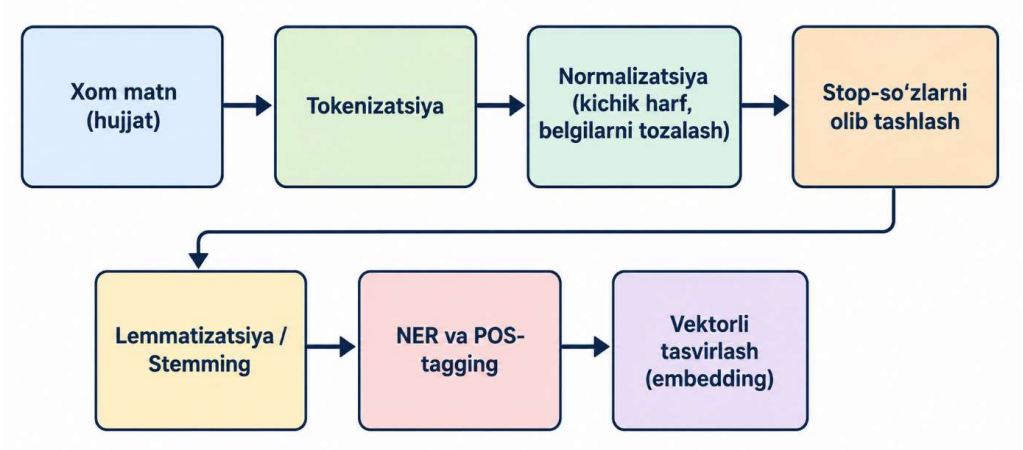
2-jadval.

Matnli ma’lumotlarni sun’iy intellekt asosida qayta ishlash usullarining qiyosiy tahlili

Usul yoki algoritm	Ishlash tamoyili	Afzalliklari	Kamchiliklari
TF-IDF	So‘z chastotasiga asoslangan statistik og‘irlik	Sodda, tez, yengil hisoblash	Semantik bog‘liqlikni hisobga olmaydi
Word2Vec / GloVe	So‘zlarni past o‘lchamli vektor fazoga joylashtirish	So‘zlar orasidagi o‘xshashlikni ifodalaydi	Kontekstni (polisemiyani) hisobga olmaydi

RNN / LSTM	Ketma-ketlikni rekurrent tarzda qayta ishlash	Kontekstual bog‘liqlikni qisman saqlaydi	Uzoq ketma-ketliklarda gradient yo‘qolishi, sekin o‘qitish
Transformer (Attention)	O‘z-o‘ziga e’tibor (self-attention) mexanizmi	Uzoq masofali bog‘liqliklarni yaxshi ushlaydi, parallellashtiriladi	Katta hisoblash resursi va ma’lumot talab etadi
BERT va uning variantlari	Ikki yo‘nalishli (bidirectional) oldindan o‘qitilgan til modeli	Yuqori aniqlik, moslashtirish (fine-tuning) imkoniyati	Model hajmi katta, mahalliy tilga moslash zarur

Transformer arxitekturasi va unga asoslangan BERT modellari eng yuqori aniqlikni ta’minlaydi, biroq bu katta hisoblash resursi va model hajmi hisobiga erishiladi. Matnni qayta ishlashning to‘liq bosqichlari (xom matndan boshlab vektorli tasvirlashgacha) 1-rasmda sxematik tarzda ifodalangan.



1-rasm. Sun’iy intellekt asosida matnli hujjatlarni qayta ishlash bosqichlari.

3-jadvalda embedding modellari va vektorli ma’lumotlar bazalarining qiyosiy xususiyatlari keltirilgan. Natijalarga ko‘ra, SBERT va multilingual-E5 kabi modellar mos ravishda 384-768 va 768-1024 o‘lchamli vektorlar generatsiya qiladi, OpenAI text-embedding xizmati esa 1536 o‘lchamli vektordan foydalanadi. Vektorlarni saqlash va qidirish uchun FAISS, Milvus va Qdrant kabi vositalar mos ravishda lokal yuqori tezlikdagi qidiruv, katta hajmdagi taqsimlangan arxiv hamda metama’lumot bilan gibrid qidiruv vazifalarini bajaradi.

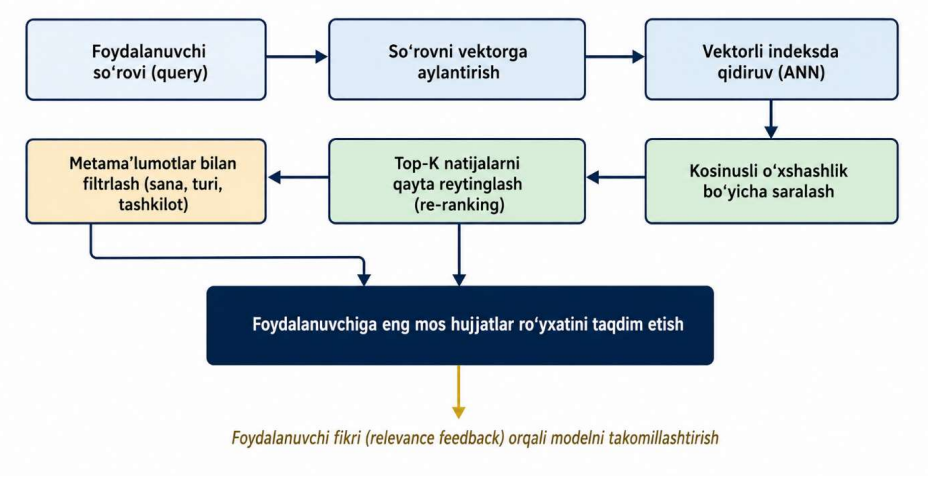
3-jadval.

Embedding modellari va vektorli ma’lumotlar bazalarining qiyosiy tahlili

Model yoki vosita	Turi	O‘lcham (vektor)	Qo‘llanish sohasi
Sentence-BERT (SBERT)	Jumla darajasidagi embedding	384-768 o‘lcham	Semantik o‘xshashlik, qidiruv
multilingual-E5	Ko‘p tilli matn embeddingi	768-1024 o‘lcham	Ko‘p tilli (jumladan o‘zbek tili) qidiruv
OpenAI text-embedding	Bulutli API embedding xizmati	1536 o‘lcham	Umumiy maqsadli semantik qidiruv
FAISS	Vektorli indekslash	-	Lokal, yuqori

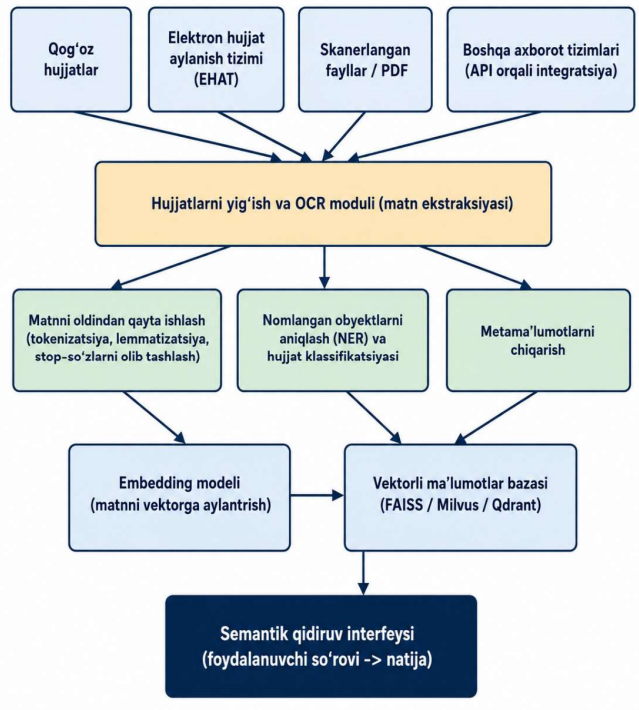
	kutubxonasi		tezlikdagi o‘xshashlik qidiruvi
Milvus	Taqsimlangan vektorli MB	-	Katta hajmdagi hujjatlar arxivi
Qdrant	Vektorli MB + filtrlash	-	Metama'lumot bilan gibrid qidiruv

Semantik qidiruv jarayonining to‘liq bosqichlari bu so‘rovni vektorga aylantirishdan tortib, natijalarni reytinglash va metama’lumot bo‘yicha filtrlashgacha 2-rasmda keltirilgan.



2-rasm. Semantik qidiruv jarayonining bosqichma-bosqich modeli.

Yuqoridagi uch yo‘nalish bo‘yicha olingan natijalar asosida davlat tashkiloti hujjatlarini sun’iy intellekt asosida qayta ishlash va semantik qidiruv tizimining besh qatlamli kontseptual arxitekturasi shakllantirildi (3-rasm).



3-rasm. Davlat tashkiloti hujjatlarini SI asosida qayta ishlash tizimining umumlashtirilgan arxitekturasi.

Arxitektura quyidagi qatlamlardan iborat:
1) Hujjat manbalari qatlami;
2) OCR va matn ekstraksiyasi qatlami;
3) NLP qayta ishlash qatlami;
4) Embedding va vektorli saqlash qatlami;
5) Semantik qidiruv interfeysi qatlami.

MUHOKAMA

Tahlil natijalari kirish qismida qo‘yilgan maqsad va vazifalarga mos ravishda talqin etilishi mumkin. Birinchi vazifa doirasida aniqlanishicha, davlat tashkilotlarida qo‘llaniladigan elektron hujjat aylanish tizimlari mazmuniy qidiruv emas, balki, birinchi navbatda, hujjat maqomini nazorat qilishga xizmat qiladi (1-jadval), bu esa kirish qismida ko‘tarilgan muammoni ya’ni an’anaviy tizimlarda semantik qidiruv imkoniyatining yo‘qligini tasdiqlaydi.

Ushbu xulosa adabiyotlar sharhida keltirilgan tadqiqotlar natijalari bilan mos keladi ya’ni milliy tadqiqotchilar ham amaldagi EHAT tizimlarida mazmuniy tahlil funksiyalarining rivojlantirilmaganligini ta'kidlaganlar [3]. Shu bilan birga, xalqaro tijorat yechimlarida SI integratsiyasining mavjudligi (1-jadval) shuni ko‘rsatadiki, texnologik jihatdan semantik qidiruvni EHAT tizimlariga qo‘shish printsipial jihatdan mumkin, muammo ko‘proq mahalliylashtirish va narx-imkoniyat nisbatida yotadi.

Matnli ma’lumotlarni qayta ishlash usullari bo‘yicha olingan natijalar (2-jadval) transformer arxitekturasining statistik va rekurrent usullardan ustunligini yana bir bor tasdiqlaydi, bu tendensiya adabiyotlar sharhida ham qayd etilgan edi [4, 5]. Ayni paytda, aniqlik va hisoblash resursi o‘rtasidagi murosa davlat tashkilotlari sharoitida alohida ahamiyat kasb etadi: cheklangan IT-infratuzilmaga ega muassasalar uchun yengil, biroq yetarlicha aniq modellarni (masalan, distillangan yoki siqilgan transformer versiyalarini) tanlash maqsadga muvofiq bo‘lishi mumkin.

Semantik qidiruv bo‘yicha natijalar (3-jadval, 2-rasm) ko‘p tilli embedding modellarining, xususan multilingual-E5 kabi yechimlarning, davlat tili siyosati nuqtai nazaridan alohida ahamiyatga ega ekanligini ko‘rsatadi, chunki ular o‘zbek tilidagi hujjatlarni qo‘shimcha katta hajmdagi maxsus o‘qitishsiz ham qoniqarli darajada qayta ishlash imkonini beradi [12]. Shu bilan birga, RAG yondashuvining joriy etilishi [17] davlat xizmati sifatini oshirishi mumkin bo‘lsada, generativ javoblarning ishonchliligini nazorat qilish mexanizmisiz bunday yondashuvni davlat sektorida qo‘llash noaniq yoki noto‘g‘ri ma'lumot berish xavfini keltirib chiqarishi mumkin, bu masala alohida e’tiborni talab qiladi.

Tadqiqotning bir qator cheklovlarini alohida ta’kidlash lozim. Birinchidan, taklif etilgan arxitektura hozircha faqat nazariy modellashtirish darajasida ishlab chiqilgan bo‘lib, uning amaliy samaradorligi maxsus dasturiy vosita yaratish va uni sinovdan o‘tkazish orqali tasdiqlanishi lozim. Ikkinchidan, tahlilga kiritilgan manbalar asosan ochiq kirish imkoniyatiga ega bo‘lgan ilmiy va me’yoriy-huquqiy hujjatlar bilan cheklangan, bu esa ba’zi tijorat tizimlarining to‘liq texnik xususiyatlarini hisobga olishni murakkablashtirgan. Uchinchidan, o‘zbek tilidagi maxsus hujjat korpuslari va etalon (benchmark) ma’lumotlar to‘plamining hozircha yetarlicha rivojlanmaganligi

embedding modellarini mahalliy tilga moslashtirish jarayonini murakkablashtiradi.

Shunga qaramay, olingan natijalar davlat boshqaruvida raqamli transformatsiyaning navbatdagi bosqichi, hujjatlar bilan ishlashni nafaqat tezlashtirish, balki ularni mazmunan “tushunadigan” tizimlarga o‘tish zaruriyatini yaqqol ko‘rsatadi. Bu o‘z navbatida, fuqarolarga ko‘rsatiladigan davlat xizmatlari sifatini oshirish, boshqaruv qarorlarini qabul qilish tezligini oshirish hamda idoralararo axborot almashinuvini soddalashtirish imkonini beradi.

XULOSA

O‘tkazilgan tahlil davlat tashkilotlarida hujjat aylanishini takomillashtirishning zamonaviy yo‘nalishi sun’iy intellekt, xususan, tabiiy tilni qayta ishlash va semantik qidiruv texnologiyalarini chuqur integratsiyalashdan iborat ekanligini ko‘rsatdi. Tadqiqot natijasida kirish qismida qo‘yilgan barcha vazifalar izchil bajarildi ya’ni elektron hujjat aylanish tizimlarining xususiyatlari, matnli ma’lumotlarni sun’iy intellekt asosida qayta ishlash usullari hamda semantik qidiruv, embedding va vektorli ma’lumotlar bazalarining zamonaviy yondashuvlari tizimli tahlil qilindi.

Tadqiqot natijasida shakllantirilgan besh qatlamli kontseptual arxitektura davlat tashkilotlari uchun milliy tilga moslashtirilgan, ko‘p tilli va kengaytiriladigan semantik qidiruv tizimini yaratishning nazariy asosi bo‘lib xizmat qiladi. Ishning amaliy ahamiyati shundaki, taklif etilgan model davlat organlarida hujjatlarni qayta ishlash va qidirish jarayonlarini avtomatlashtirish bo‘yicha dasturiy vositalarni loyihalashda bevosita foydalanilishi mumkin.

Kelgusi tadqiqotlar doirasida ushbu arxitekturaning amaliy dasturiy vositasi ishlab chiqiladi, o‘zbek tilidagi hujjatlar korpusi asosida embedding modeli moslashtiriladi (fine-tuning) hamda tizimning qidiruv aniqligi va tezligi bo‘yicha eksperimental baholash o‘tkaziladi.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI.

1. O‘zbekiston Respublikasi Prezidentining “Raqamli O‘zbekiston - 2030” strategiyasini tasdiqlash to‘g‘risida”gi Farmoni, 2020-yil 5-oktabr, PF-6079-son.
2. O‘zbekiston Respublikasi Vazirlar Mahkamasining elektron hujjat aylanish tizimlarini joriy etish to‘g‘risidagi qarori. - Toshkent, 2021.
3. Karimov B.R., Yusupov T.A. Davlat boshqaruvida raqamli texnologiyalar: elektron hujjat aylanishi tahlili. O‘zbekiston axborot texnologiyalari jurnali. - 2022. - №3. - B. 45-52.
4. Vaswani A., Shazeer N., Parmar N. va boshq. Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS). - 2017. - P. 5998-6008.
5. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT. - 2019. - P. 4171- 4186.
6. ISO 15489-1:2016. Information and documentation - Records management - Part 1: Concepts and principles. - Geneva: International Organization for Standardization, 2016.

7. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781. - 2013.
8. Brown T., Mann B., Ryder N. va boshq. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems (NeurIPS). - 2020. - P. 1877-1901.
9. Toshpulatov D.A. O‘zbek tilidagi matnlarni avtomatik qayta ishlashning lingvistik asoslari. Til va adabiyot ta’limi jurnali. - 2021. - №2. - B. 12-19.
10. Bishop C.M. Pattern Recognition and Machine Learning. - New York: Springer, 2006. - 738 p.
11. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP-IJCNLP. - 2019. - P. 3982-3992.
12. Wang L., Yang N., Huang X. va boshq. Multilingual E5 Text Embeddings: A Technical Report. arXiv preprint arXiv:2402.05672. - 2024.
13. Johnson J., Douze M., Jegou H. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data. - 2021. - Vol. 7, No. 3. - P. 535-547.
14. Malkov Y.A., Yashunin D.A. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence. - 2020. - Vol. 42, No. 4. - P. 824-836.
15. Manning C.D., Raghavan P., Schutze H. Introduction to Information Retrieval. - Cambridge: Cambridge University Press, 2008. - 544 p.
16. Salton G., McGill M.J. Introduction to Modern Information Retrieval. - New York: McGraw-Hill, 1983. - 448 p.
17. Lewis P., Perez E., Piktus A. and others. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems (NeurIPS). - 2020. - P. 9459-9474.
18. Douze M., Guzhva A., Deng C. and others. The Faiss library. arXiv preprint arXiv:2401.08281. - 2024.