UO‘K 004.8

O‘ZBEK TILI UCHUN KENGAYTIRILGAN KONTEKSTLI MODERNUZBERT SEMANTIK EMBEDDING MODELINI ISHLAB CHIQISH VA SAMARADORLIGINI BAHOLASH

Xujayarov I.Sh.[1], Ochilov M.M.[1], [+] Xolmatov O.A.[1], Jumanov V.I.[2]

[1]Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti, Toshkent, O‘zbekiston

[2] “Adliya organlari va muassasalarida axborot-kommunikatsiya texnologiyalarini rivojlantirish markazi” davlat muassasasi, Toshkent, O‘zbekiston

[+] xolmatov.orzumurod@gmail.com

Annotatsiya. Tadqiqot o‘zbek tili uchun zamonaviy transformer arxitekturasiga asoslangan, kengaytirilgan kontekst oynasiga (8192 token) ega ModernUzBERT embedding modelini ishlab chiqishga bag‘ishlangan. O‘zbek tili kabi kam resursli va agglyutinativ tillarda keng qamrovli kontekstni qayta ishlash imkoniyati cheklanganligi model samaradorligiga salbiy ta’sir ko‘rsatadi. Ushbu muammoni hal etish uchun 125 milliondan ortiq so‘z shaklini o‘z ichiga olgan o‘zbek lotin yozuvidagi yirik matnli korpus shakllantirildi va 52 000 sub-word birligidan iborat optimallashtirilgan lug‘at yaratildi. Modelni o‘qitish jarayoni ikki bosqichda amalga oshirildi: dastlab model Masked Language Modeling (MLM) usulida noldan o‘qitildi, so‘ngra 30 000 ta savol-javob juftligidan iborat dataset yordamida Supervised Fine-Tuning (SFT) orqali semantik qidiruvga moslashtirildi. Flash Attention 2 va Unpadding texnologiyalarining qo‘llanilishi hisobiga GPU resurslaridan foydalanish samaradorligi sezilarli darajada oshdi. 8192 tokenli kontekst oynasi katta hajmli matnlarni semantik yaxlitligini saqlagan holda tahlil qilish imkonini beradi. Tadqiqot yakunida ishlab chiqilgan model va uning barcha komponentlari ilmiy hamjamiyat uchun Hugging Face platformasida ochiq foydalanishga taqdim etildi.

Kalit so‘zlar: NLP, o‘zbek tili, ModernBERT, semantik embedding, axborot qidiruvi, katta o‘lchamli kontext, flash attention.

1 KIRISH

Tabiiy tilni qayta ishlash (NLP) sohasida transformer arxitekturasiga asoslangan modellar tufayli ko‘plab yutuqlarga erishilmoqda. Ammo o‘zbek tili kabi kam resursli tillar uchun yuqori aniqlikdagi embedding modellarini yaratish hamon muammo sanaladi. Semantik qidiruv, savol-javob tizimlari va axborotni qayta ishlash texnologiyalari nafaqat sintaktik tuzilishini, balki ushbu matnning umumiy semantik ma’nosini tushunishni ham talab qiladi. Aynan o‘zbek tili uchun NLP sohasida embedding modellar cheklanganligi sababli raqamli iqtisodiyot, yuridik, ta’lim kabi sohalarda matnlarni avtomatik tahlil qilish rivojlanishi uchun katta to‘siq bo‘lmoqda.

Hozirgi vaqtda o‘zbek tilini qo‘llab-quvvatlaydigan BGE va uzBERT kabi ko‘plab modellar katta hajmdagi matnlarni embedding shaklida ifodalashda tokenlar soni, lug‘at o‘lchami va hisoblash samaradorligi bilan bog‘liq qator cheklovlarga ega. Xususan, hozirgi kundagi standart modellar ko'pi bilan 512 tokengacha bo'lgan matnlarni qayta ishlay oladi. Bu esa yirik hujjatlar, ilmiy maqolalar yoki hisobotlarni yaxlit embedding shaklida tahlil qilishda tokenlar chegarasi bilan bog‘liq muammolarni keltirib chiqaradi.

O‘zbek tili agglyutinativ til sifatida boy so‘z yasovchi qo‘shimchalarga ega. Hozirgi kunda mavjud modellarning lug‘at qamrovi nisbatan kichik bo‘lib, ko‘p miqdordagi unikal so‘z shakllarini o‘zida jamlagan murakkab morfologik strukturani to‘liq qamrab ololmaydi. Natijada, lug‘atdan tashqari so‘zlar (OOV – Out-Of-Vocabulary) muammosi yuzaga keladi [1].

An'anaviy arxitekturalar katta hajmdagi matnlarni qayta ishlashda GPU resurslaridan samarali foydalanish imkoniyatini bermaydi, bu esa real vaqt rejimida ishlashda kechikishlarga sabab bo‘lishi mumkin.

Ushbu maqolada zamonaviy ModernBERT arxitekturasini o‘zbek tilining lingvistik xususiyatlariga moslashtirish orqali yuqori samaradorlikka ega semantik modelni ishlab chiqish rejalashtirilgan. Bunda modelning kontekst oynasini 8192 tokengacha kengaytirish orqali, katta hajmdagi matnlarni o‘zining

semantik yaxlitligini yo‘qotmagan holda chuqur qayta ishlash imkonini taqdim etish. Shuningdek, o‘zbek tilining lotin grafikasiga asoslangan murakkab morfologiyasini to‘laqonli qamrab olish maqsadida, 10 milliondan ortiq jumlalar korpusi asosida 1 348 641 ta unikal so‘zdan iborat maxsus lug‘at bazasi shakllantirish hisobiga samarli yechimga erishish ta’minlandi. Shu bilan bir qatorda, tadqiqotda Flash Attention va Rotary Positional Embedding (RoPE) kabi ilg‘or mexanizmlarni qo‘llash orqali arxitekturani optimallashtirish hamda modelning aniqlik darajasi va hisoblash tezligini sezilarli darajada oshirish nazarda tutilgan.

2 ADABIYOTLAR TAHLILI

O‘zbek tili uchun NLP sohasida so‘ngi tillarda ko‘plab modellar ishlab chiqilgan. Dastlabki bosqichda Google tomonidan ishlab chiqilgan mBERT (multilingual BERT) va META kompaniyasining XML-RoBERTA (XLM-R) kabi ko‘p tilli modellar o‘zbek tili uchun asosiy vosita hisoblandi [2]. Ammo, ushbu modellar ko‘p tillarni qo‘llab quvatlagani uchun ularning lug‘at boyligida o‘zbek tili ulushi juda kichik (odatda 1-2%) bo‘lib, bu o‘zbek tili kabi agglyutinativ tillarda murakkab morfologik strukturani to‘liq qamrab olishga muammo yuzaga keltiradi [3].

Keyinchalik o‘zbek tili uchun maxsus mono-lingual modellar paydo bo‘ldi. Jumladan, UzRoBERTA modeli o‘zbek tili korpusida o‘qitilgani uchun semantik aniqlikda ko‘p tilli modellardan oshib ketdi. Shu sababli, Adilova Fatima va boshqalar tomonidan ishlab chiqilgan UzRoBERTa modeli o‘zbek tilini qo‘llab quvatlaydigan vazifalarda asosiy poydevor vazifasini bajarmoqda [4, 5].

ModernBERT arxitekturasi zamonaviy avlod arxitekturasi sanaladi va uning an’anaviy modellardan farqi 2024-yil oxirida taqdim etilgan ModernBERT arxitekturasi BERT (Enkoder only) modellarining zamonaviy modeli hisoblanadi. U nafaqat tezlik, balki bundan tashqari ma’lumotlarni qayta ishlash sifati bilan eski avlod modellardan katta farq qiladi [6]. ModernBERTning jahon miqiyosidagi ilmiy yangiligi va an’anaviy modellardan asosiy farqi quyidagilardan iborat:

- qo‘llab quvvatlanadigan kontext uzunligi – ModernBERT “sliding window attention” va maxsus pozition embeddinglar yordamida 8192 tokengacha bo‘lgan matnlarni bitta embeddinga sig‘dira oladi. Bu esa UzRobERTadan 16 marta katta hajmdagi matni qo‘llab quvvatlashini bildiradi [7].
- Zamonaviy BERT modellarida matnlarni bir xil uzunlikga keltirish uchun “padding” ishlatiladi, bu o‘z navbatida GPU resurslarining bekorga sarflanishiga olib keladi. ModernBERT “unpadding” texnologiyasi orqali faqat samarali tokenlarni hisoblaydi, bu esa o‘z navbatida hisoblash tezligini 15-20% ga oshiradi [8, 28, 31].
- ModernBERT zamonaviy GPU (NVIDIA A100/H100) arxitekturalari uchun maxsus optimallashtirilgan bo‘lib, xorita sarfini (VRAM) sezilarli darajada kamaytiradi.
- GeLU dan GeGLU ga o‘tish orqali modelning ichki aktivatsiya funksiyalari o‘zgartirilishi sababli semantik ma’nolarni o‘rganish samaradorligi oshgan [9, 24].

So‘nggi yillarda NLP sohasida transformer arxitekturasi asosidagi modellar, xususan BERT, savol-javob tizimlarini yaratishda keng qo‘llanilmoqda. Biroq klassik BERT modeli kontekst uzunligining cheklanganligi va hisoblash samaradorligi bilan bog‘liq muammolar tufayli murakkab hujjatlar bilan ishlashda yetarli darajada samarali emas.

Ushbu muammolarni hal etish maqsadida AnswerDotAI jamoasi tomonidan taklif etilgan **ModernBERT** modeli zamonaviy optimizatsiya usullarini o‘zida mujassam etgan yangi avlod encoder modeli hisoblanadi. Alexis va boshqalar “Unsupervised Cross-lingual Representation Learning at Scale” nomli maqolasida model 2 trillion token ustida o‘qitilgani va 8192 tokenli kontekst uzunligini qo‘llab-quvvatlashi qayd etilgan. Mualliflar tomonidan ta’kidlanishicha, ushbu model GLUE va boshqa benchmarklarda klassik BERTga nisbatan yuqori natijalarni ko‘rsatadi [10].

ModernBERT arxitekturasining asosiy komponentlaridan biri bo‘lgan Rotary Position Embedding (RoPE) mexanizmi Su va boshqalar tomonidan taklif etgan RoPE mexanizmi tokenlar o‘rtasidagi nisbiy pozitsion bog‘lanishni yaxshiroq ifodalashi va uzun kontekstlarda samarali ishlashi keltirilgan [11]. Shuning bilan birgalikda ModernBERT modelida qo‘llanilgan FlashAttention mexanizmi hisoblash samaradorligini oshirishda muhim ahamiyat kasb etishini, Dao va boshqalar keltirib o‘tganlar [12]. Shuningdek, Zaheer uning jamoasi “sparse attention” mexanizmi yordamida uzun ketma-ketliklarni qayta ishlash usulini taklif etganlar [13].

ModernBERT modeli turli sohalarda, jumladan tibbiyotda ham samarali natija ko‘rsatishi hiqida Zhang va boshqalar modelni klinik matnlar uchun moslashtirganligini, named entity recognition va matn klassifikatsiyasi vazifalarida yuqori natijalarni ko‘rsatganligi bildirgan [14]. Shuningdek, Wang va boshqalar ishida ModernBERT modeli 53 milliard token asosida o‘qitilib, biotibbiy NLP vazifalarida samarali ishlashi ko‘rsatgan [15]. Turli tillarda ModernBERT modellarini ishlab chiqish bo‘yicha ham izlanishlar mavjud. Jumladan, Tanaka va boshqalar tomonidan ModernBERT modelini yapon tiliga moslashtirish taklif etilgan [16].

Savol-javob tizimlarida hujjatlar orasidan mos javobni topish uchun retrieval mexanizmlari muhim ahamiyatga ega. Shu nuqtai nazardan, Li va boshqalar o‘z tadqiqot ishida ModernBERT modelini encoder sifatida qo‘llab, ColBERT modeli bilan integratsiya qilgan va retrieval aniqligi oshirilgani ko‘rsatilgan.[17] Bundan tashqari, ModernBERT multimodal tizimlarda ham qo‘llanilmoqda. Jumladan, Chen va boshqalar tomonidan ModernBERT modeli asosida matn va tasvirni birlashtirgan model taklif etilgan [18].

Yuqoridagi tahlillardan ko‘rinadiki, ModernBERT modeli savol-javob tizimlari uchun muhim bo‘lgan quyidagi afzalliklarga ega:

- uzun kontekst bilan ishlash imkoniyati
- yuqori hisoblash samaradorligi
- retrieval va RAG tizimlari bilan integratsiya qilish imkoniyati

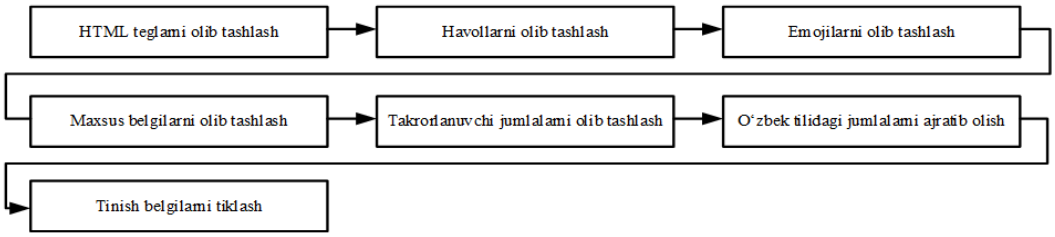
Shu bilan birga, mavjud tadqiqotlar asosan ingliz va yuqori resursli tillarga qaratilgan bo‘lib, o‘zbek tili uchun ModernBERT asosidagi savol-javob tizimlari deyarli mavjud emas. Bu esa mazkur yo‘nalishda yangi ilmiy izlanishlar olib borish uchun muhim ilmiy bo‘shliq mavjudligini ko‘rsatadi.

3 METODOLOGIYA

Ma’lumotlar to‘plami va ko‘rpusni shakllantirish. Korpus turli sohalarni o‘z ichiga olgan bo‘lib, bu modelning o‘zbek tilini bu modelning semantik tushinish qobiliyatini yuqori cho‘qqiga olib chiqishini ta’minlaydi. Ushbu ma’lumotlar quyidagilardan jamlangan:

- Rasmiy-huquqiy: O‘zbekiston Respublikasi qonun xujjatlari ma’lumotlar bazasi (lex.uz);
- Ensiklopedik: Wikipedia o‘zbek bo‘limining to‘liq arxivi;
- Publitsistik: Yirik milliy yangiliklar agregatorlari va ommaviy axborot vositalari;
- Badiiy: Raqamlashtirilgan o‘zbek mumtoz va zamonaviy adabiyoti namunalari.

Tozalash va qayta ishlash. Yig‘ilgan ma’lumotlarni samaradorligini oshirish uchun ko‘p bosqichli quyidagi tozalash ishlari 1-rasmda keltirilgani kabi amalga oshirildi:



1-rasm. Yig‘ilgan matnli ma’lumotlarni tozalash bosqichlari

Yig‘ilgan korpus qayta ishlash bosqichidan keyin quyidagi 1-jadvalda keltirilgan parametrlarga ega korpus paydo bo‘ldi.

1-jadval. Tadqiqot uchun shakllantirilgan korpus parametrlari

Ko‘rsatkich	Miqdor/tavsifi
Unikal jumlalar soni	10 000 000 +
Jami so‘zlar soni	125 261 608
Lug‘at hajmi(unikal so‘zlar)	1 348 641
Til	O‘zbek lotin yozuvi

Ushbu jarayondan keyin ishlab chiqilgan ko‘rpusdan foydalanib, 40 000 ta kontext, savol, javob ustunlaridan iborat dataset hosil qilindi. O‘zbek tili agglyutinativ tuzilishiga ega bo‘lgani uchun bir o‘zakdan ko‘plab so‘z shaklini hosil qilishi mumkin. Ushbu holatni hisobga olgan holda Byte-Pair encodeingn (BPE) modelidan foydalanildi [19].

Mavjud modellardan farqli ravishta lug‘at hajmi 1 348 641 ta unikal so‘z shakllaridan iborat. Bu o‘zbek tilidagi kam uchraydigan terminlar va murakkab affikslarda OOV muammosini bartaraf etish imkonini beradi.

O‘qitish jarayoni 2-jadvalda ko‘rsatilgan 2 ta asosiy bosqichdan iborat bo‘lib, u pre-training va fine tuning jarayonlarini o‘z ichiga oladi.

Dastlabki bosqichda model o‘zbek tilining fundamental sintaktik va semantik qonuniyatlarini o‘rganish maqasadida Masked language model (MLM) texnikasidan foydalanildi. Bunda matn tarkibidagi tokenlarning 15% qismi tasodifiy shaklda niqoblanib (mask), model kontextga qarab ushbu tokenlarni bashorat qilishga o‘rgatildi.

2-bosqichda MLM bosqichidan keyin modelni qidiruv va savol javob vazifalariga moslashtirish uchun tayyorlangan 40 000 ta context, savol, javob ustunga ega bo‘lgan datasetning 30 000 ta qatordan iborat qismi ajratib olindi. 10 000 qatorli qismi esa keyinchalik baholash uchun qoldirildi. Ushbu bosqichda Constractive learning (qarama-qarshi o‘qitish) yondashuvi va multiple Negatives Ranking loss funksiyasi qo‘llanildi. Bundan maqsad modelga nafaqat so‘zlar ketma-ketligini balki savol va unga mos javobni semantik fazoda bir biriga yaqin joylashish kerakligini ta’minlash.

2-jadval. ModernUzBERTmodelini o‘qitishda foydalanilgan parametrlar

Parametr nomi	1-bosqich. MLM (Pre-train)	2-bosqich. SFT(Fine-tune)
Ma’lumotlar hajmi (so‘z)	125 000 000 +	30 000 (savol- javob juftligi)
Learning rate	$2* 10^{-5}$	$1*10^{-5}$
Batch size	64	32
Epoch	5	3

Tadqiqot doirasida 8192 tokenli kontekst oynasi va 128 million tokenli ma'lumotlar to'plamiga ega Modern BERT modelini o‘qitish hamda aprobatsiyadan o‘tkazish maqsadida optimallashtirilgan yuqori unumdor hisoblash infratuzilmasidan foydalanildi.

Katta hajmli matnli korpuslarni yuklash, tokenizatsiya qilish va uzluksiz qayta ishlash barqarorligini ta'minlash uchun tizim 16 GB videoxotiraga ega GPU hamda 128 GB tezkor xotira asosida qurildi. Hisoblash samaradorligini oshirish va apparat resurslaridan oqilona foydalanish maqsadida diqqat mexanizmi matritsalarini hisoblashda yuzaga keladigan xotira tanqisligini bartaraf etuvchi FlashAttention-2 texnologiyasi, shuningdek, videoxotira sarfini 50% gacha tejab, hisoblash tezligini sezilarli darajada oshiruvchi BF16 (Bfloat16) aralash aniqlikdagi o'qitish (mixed precision training) usuli joriy etildi.

Natijada, ushbu apparat-dasturiy integratsiya yordamida murakkab arxitekturali til modelini o‘qitish jarayonida uzluksizlik, yuqori hisoblash unumdorligi hamda resurslar tejamkorligiga erishildi. O‘qitish jarayonida gradient chekpoint usulidan foydalanildi, o‘z o‘rnida 16 GB videoxotirali cheklovga ega bo‘lgan qurilmani, eng yuqori 8192 token uzunlikdagi ketma-ketlikni xatolarsiz ishlash imkoniyatini berdi.

4 TADQIQOT NATIJALARI

Baholash metrikalari. Ishlab chiqilgan modelning semantik aniqligini holis baholash maqsadida 10 000 ta savol, javob juftligidan iborat maxsus to‘plamdan foydalanildi. Tadqiqot “Dense Retrieval” (zich qidiruv) tamoili asosida o‘tkazildi. Model matn shaklidagi so‘rovni vektor shaklida ifodalab, katta hajmli bazadan unga eng yaqin bo‘lgan javoblarni topish kerak bo‘ladi. Buning uchun Recall@k va MRR (Mean Reciprocal Rank) baholash metrikalari qo‘llanildi [20][21].

Recall@k top k ta topilgan matn parchalari ichida savolga javob matn parchasi mavjudligini aniqlash orqali 1- ifoda orqali hisoblanadi:

$$Recall@k = \frac{Top\ k\ ichida\ tog‘ri\ matn\ parchasi\ mavjud\ bo‘lgan\ savollar\ soni}{Jami\ savollar\ soni}. \quad (1)$$

Tadqiqotda berilgan savolga mos keluvchi matn parchasini aniqlash samaradorligini hisoblashda Recall@1, Recall@3, Recall@5, Recall@10 qiymatlaridan foydalanildi.

MRR baholash metrikasi qidiruv natijalaridagi to‘g‘ri matn parchasining o‘rnini hisobga olgan holda quyidagi 2-ifoda orqali aniqlanadi:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}, \quad (2)$$

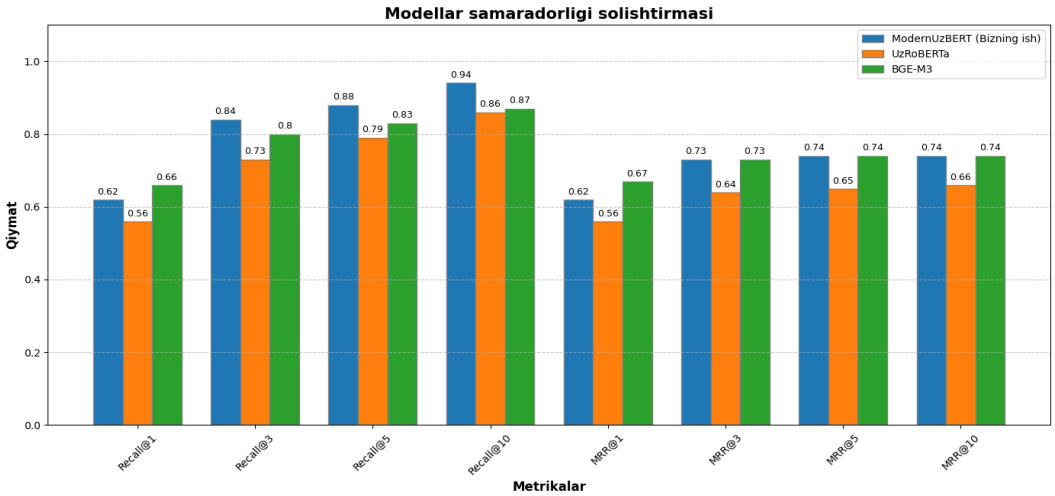
bu yerda N savollar soni, ranki i savol uchun tog‘ri kontextning natijalar ro‘yxatidagi o‘rnini bildiradi. MRR qiymati qanchalik yuqori bo‘lsa, tizim shunchalik samarali ishlayotganligini tushunish mumkin.

Muqobil modellar va ModernUzBERT modelining qiyosiy tahlili. ModernUzBERTmodel hozirgi kunda o‘zbek tili uchun NLP vazifalarida muhim poydevor bo‘lgan UzRoBERTA va ko‘p tilli BGE-M3 global modellari bilan solishtirildi. Solishtirish orqali 3-jadvalda keltirilgan natijalar olindi. Jadval natijalarini 2-rasm kabi diagramma shaklida tasvirlansa yaqqol ustun bo‘lgan model ko‘zga tashlanadi.

Olib borilgan tahlillar shunin ko‘rsatadiki,ishlab chiqilgan ModernUzBERT modeli o‘zbek lotin yozuvidagi semantik qidiruv vazifalarini bajarishda yuqori natijalarni ko‘rsatadi. Tadqiqotda Recall@k va MRR@j baholash metrikalari bo‘yicha nafaqat bazaviy UzRoBERTa modeli bilan balki global ko‘p tilli modellar (jumladan BGE-M3) modellar bilan raqobatlasha olishini ko‘rsatdi.

3-jadval. Modellarning semantik qidiruv samaradorligi

Metrika	ModernUzBERT (Ushbu ish)	UzRoBERTa	BGE-M3
Recall@1	0.62	0.56	0.66
Recall@3	0.84	0.73	0.80
Recall@5	0.88	0.79	0.83
Recall@10	0.94	0.86	0.87
MRR@1	0.62	0.56	0.67
MRR@3	0.73	0.64	0.73
MRR@5	0.74	0.65	0.74
MRR@10	0.74	0.66	0.74



2-rasm. O‘zbek tili embeding modellarining semantik qidiruv samaradorligi

ModernUzBERT va BGE-M3 modellarini solishtirish. Olib borilgan tadqiqotga natijalariga ko‘ra, jumladan recall@10 ko‘rsatkichi 0.94 ga yetgan bo‘lsa, aynan ushbu ko‘rsatgich BGE-M3 modelida 0.87 va UzRoBERTa modelida esa 0.86 ni ko‘rsatdi. Bu esa ModernUzBERT modeli qidiruv qamrovida boshqa muqobil modellardan sezilarli darajada o‘zib ketganligini ko‘rsatadi. Recall@1 baholash ko‘rsatgichi bo‘yicha MGE-M3 modeli 0.66 va ModernUzBERT modeli 0.62 natija orqali biroz ustun ekanligini ko‘rsatgan bo‘lsada, k qiymati ortishi bilan ModernUzBERT modeli samaradorlik jihatdan ustunlik qildi. Bu esa o‘zbek tilida o‘qitilgan modelning murakkab va katta kontextdan tashkil topgan so‘rovlarda javob topish ehtimoli yuqoriligini bildiradi.

ModernUzBERT va UzRoBERTa modellarini solishtirish. Hozirgi kunda tayanch vazifasini bajaradigan UzRoberta modelidan ishlab chiqilgan ModernUzBERTmodelining afzalligi shundaki recall@10 baholash metrikasida 8% yuqoriroq natijani ko‘rsatdi.

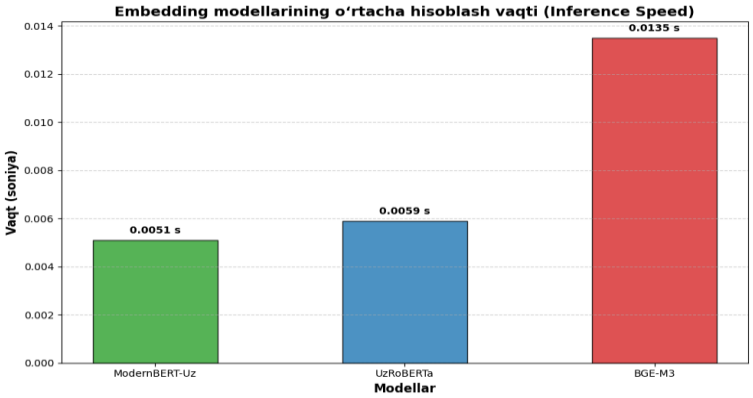
Savolga mos javoblarning tartibi bo‘yicha samaradorlik (MRR@K). MRR baholash metrikasi model taqdim etilgan to‘g‘ri javob nechanchi o‘rinda ekanligini baholaydi. Ushbu baholash metrikasida ModernUzBERTmodeli MRR@10 ko‘rsatkichi bo‘yicha 0.74 qiymat bilan BGE-M3 modeli bilan teng natija ko‘rsatgan bo‘lsa, UzRoBERTa modeli 0.66 natijani ko‘rsatdi. Bu esa UzRoBERTa modelidan ModernUzBERT modeli sezilarli darajada yuqori ekanligini ko‘rsatdi. Ushbu natija esa ModernUzBERT modeli nafaqat javobni topishi, balki shuning bilan birga qidiruv natijasida savolga mos ravishta javoblarni eng yuqori qismida (Top 3 ichida) joylashtirish imkoniyatini ko‘rsatdi.

Barqarorlik va arxitektura tomonidan ustunlik. Tadqiqot natijalari shuni ko‘rsatadiki UzRoBERTa modeli 512 tokendan uzun bo‘lgan matnlarni qayta ishlashda “index out of bounds” xatosiga uchraydi, shu sababli ushbu model katta o‘lchamdagi matnlar bilan ishlashda muamolarga uchrashini ko‘rsatadi. ModernUzBERT modeli esa 8192 token qo‘llab quvatlay olishi va Rotary Positional Embeddings (RoPE) mexanizmidan foydalanilganligi sababli, matni bo‘laklamasdan to‘laligicha semantik yaxlitligini saqlagan holda tahlil qilish imkoniyatini beradi [22,23].

Hisoblash tezligi. Zamonaviy NLP vazifalarida ayniqsa RAG va real vaqtda ishlovchi chatbotlar uchun modelning kechikish vaqti muhim hisoblanadi. Shu sababli bitta so‘rovga embedding yaratish uchun sarflaydigan o‘rtacha vaqti hisoblandi va ushbu natijalar 4-jadvalda aks ettirildi. 4-jadval natijalarnidan foydalanib diagramma hosil qilinsa 3-ramda tasvirlangan tezlik jihatdan eng tezkor modelni ko‘rish mumkin.

4-jadval. Modellarning o‘rtacha hisoblash tezligi (soniyalarda)

Model nomi	O‘rtacha vaqt(s)	Samaradorlik (%)
ModernBERT-Uz	0.0051	164.7% tezroq
UzRoBERTa	0.0059	128.8% tezroq
BGE-M3	0.0135	0.0 %



3-rasm. Matnlarni vektorlashtirish bo‘yicha modellar solishtirmasi

Tadqiqotda foydalanilgan modellarning so‘rovlarni qayta ishlash tezligi bo‘yicha tadqiqotlar shuni ko‘rsatadiki, ModernUzBERT modeli tadqiqotda foydalanilgan boshqa modellardan samarali ekanligini ko‘rsatdi. Jumladan ModernUzBERT modeli bitta so‘rovni embedding shaklida tasvirlash uchun o‘rtacha 0.0051 soniya sarflagan, bu ko‘rsatgich BGE-M3 modeliga qaraganida (0.135 soniya) 2.6 marta tezroq ekanligini ko‘rsatdi. Bu natijaga Flash Atttention va optimallashtirish RoPE mehanizmlaridan foydalanish evaziga erishilgan. UzRoBERTa modeli bilan solishtirilganda ham ModernUzBERT modeli 15.6% tezroq ishlashini nomoyon etdi. Bu esa ushbu arxitekturaning nafaqat aniqlikda balki real vaqt rejimida ishlovchi tizimlar uchun ham qulay yechim ekanligini ko‘rsatadi.

BGE-M3 modeli kechikish vaqtini ko‘p tilli murakkab lug‘at bazasi va parallel hisoblash algoritmlari (dense/sparse) bilan tushuntiriladi [28, 29].

ModernUzBERT modeli UzRoBERTa modelidan tezroq ishlashining sababi modernBERT arxitekturasining Flash Attention 2 va Unpadding texnologiyalari yordamida GPU hisoblash resurslaridan foydalanish orqali bajarilishi hisoblanadi [25].

Tadqiqot natijalarida ModernUzBERT modeli UzRoBERTa modeliga qaraganda tezkor ekanligini ko‘rsatdi. Bu natija tasodifiy emas, balki ModernUzBERT arxitekturasi Flash Attention 2 texnologiyasining muvofaqiyatli integratsiya qilish natijasidir. An’anaviy transformer arxitekturalaridan foydalanishda diqqat mexanizmlari hisoblashamallari matn uzunligining kavratik proporsionallik (O(n2))ga mutanosib ravishda o‘sib boradi. Bu esa katta o‘lchamdagi matnlarni qayta ishlashda GPU xotirasi bilan almashuv kiritish/chiqarish vaqtining keskin oshib ketishiga olib keladi. Flash attention esa ushbu hisoblashlarni GPU xotirasida bloklarga bo‘lgan holda xotira o‘tkazuvchanligini optimallashtirish imkonini beradi. Natijada, model nafaqat tezlikni oshiradi, balki katta o‘lchamdagi matnli (8192 token) ma’lumotlarni qayta ishlashda samaradorlikni oshiradi [26].

Katta o‘lchamdagi matnli ma’lumotlarni qayta ishlashda, kontexda qo‘llanilgan 8192 tokenli kontext oynasi NLP sohasida sifat jihatdan yangi bosqichga olib chiqadi. Avvalgi axsavlod modellari 512 tokenli modellarda ishlay olishi sabab, katta o‘lchamdagi matnli ma’lumotlarni qayta ishlash imkoniyati mavjud emas, buning uchun katta o‘lchamdagi matnli ma’lumotlarni kichik qismlarga ajratish lozim edi [27]. Natijada umumiy xujjat semantik yaxlitlikni va uzoq masofali mantiqiy bo‘g‘liqliklarni yo‘qolishiga olib kelar edi.

Ushbu modelni RAG(Retrieval Augmented Generation) tizimlarini qurishda qo‘llash mumkin. Kengaytirilgan kontext modelga bitta so‘rov doirasida ko‘proq ma’lumotlarni qayta ishlash imkoniyatini beradi, bu esa o‘z navbatida RAG tizimlarida generator modellari (masalan, Qween, Llama yoki GPT) tomonidan shakllantiriladigan javob sifatini oshirish uchun xizmat qiladi [30].

ModernUzBERT modeli hozirgi bosqichda Masked Language Modeling(MLM) va boshlang‘ich fine-tuning bosqichidan muvafaqiyatli o‘tdi va o‘zbek tilining fundamental semantikasini o‘zlashtirdi. Kelgusida modelning qobiliyatini oshirish maqsadida modelni tor doiraga (davlat xizmatlari, tibbiyot, moliya, huquq) sohalariga moslashtirilgan ko‘rpuslarda o‘qitish orqali uning professional terminalogiya bilan ishlash imkiniyatini oshirish va RAG texnologiyalarida qo‘llash yo‘nalishlari belgilandi.

5 XULOSA

Tadqiqot natijalari tasdiqlaydiki, ModernUzBERT modeli o‘zbek tilidagi matnli ma’lumotlarnda embeddinglar yaratishda yuqori samarali hisoblanadi. Model Recall@10 metrikasida savolga mos kontextni topishda 94.6% aniqlik qayt etib, global BGE-M3 va UzRoBERTa modellaridan sezilarli darajada o‘zib ketdi.

Ishlash tezligi jihatidan ModernUzBERT modeli o‘rtacha bitta so‘rovni qayta ishlash vaqti bo‘yicha muqobil modellardan ustunligini yaqqol ko‘rsatdi. Modelning 8192 tokenli kontext oynasi va o‘zbek tili morfologiyasiga moslashgan lug‘at boyligi murakkab matnlar bilan ishlash imkoniyatini oshiradi.

Umuman olganda ModernUzBERT modeli nafaqat semantik aniqlikni oshiradi balki hisoblash jihatdan tezligini nomoyon etib, o‘zbek lotin yozuvidagi qidiruv tizimlari va katta o‘lchamli xujjatlar bilan ishlovchi RAG platformalarida matnlarni embedding shaklida tasvirlash uchun samarali yechim bo‘lib xizmat qiladi.

ADABIYOTLAR

[1] *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I.* (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS), 5998–6008

[2] *Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the NAACL-HLT. 2019. 4171–4186

[3] *Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov*. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451

[4] *Adilova, Fatima & Davronov, Rifkat & Safarov, Ruzmat.* (2023). Uzroberta: An uzbek language pre-trained model. Universum:Technical sciences. 115. 10.32743/UniTech.2023.115.10.16028.

[5] *Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692. 2019. Pages 1–13.

[6] *Smit, B., Tworkowski, K., Yazici, A.,* & ModernBERT Team. ModernBERT: A New Frontier in Encoder-Only Transformers. arXiv preprint arXiv:2412.13663. 2024. Pages 1–24.

[7] *Dao, T., Fu, D., Ermon, S., Rudra, A., & Ré, C.* FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. Advances in Neural Information Processing Systems (NeurIPS). 2022. Pages 16344–16359

[8] *Dao, T.* FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. International Conference on Learning Representations (ICLR). 2024. Pages 1–12.

[9] *Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research. 2020. Vol. 21. Pages 1–67

[10] *Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov*. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451

[11] *Su J., Ahmed M., Lu Y., Pan S., Bo W., Liu Y.* RoFormer: Enhanced Transformer with Rotary Position Embedding // arXiv preprint arXiv:2104.09864. – 2021. – P. 1–10. URL: https://arxiv.org/abs/2104.09864

[12] *Dao T., Fu D. Y., Ermon S., Rudra A., Ré C.* FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness // Advances in Neural Information Processing Systems (NeurIPS). – 2022. – Vol. 35. – P. 16344–16359. URL: https://arxiv.org/abs/2205.14135

[13] *Zaheer M., Guruganesh G., Dubey K. A., et al.* Big Bird: Transformers for Longer Sequences // Advances in Neural Information Processing Systems (NeurIPS). – 2020. – Vol. 33. – P. 17235–17251. URL: https://arxiv.org/abs/2007.1406

[14] *Zhang R., et al.* Clinical ModernBERT: Adapting ModernBERT for Clinical NLP // arXiv preprint arXiv:2412.16480. – 2024. – P. 1–12. URL: https://arxiv.org/abs/2412.16480

[15] *Wang X., Wang Z., Ahmed M., et al.* BioClinical ModernBERT: Pre-training ModernBERT on 53 Billion Tokens for Biomedical NLP // arXiv preprint arXiv:2501.06648. – 2025. – P. 1–14. DOI: 10.48550/arXiv.2501.06648

[16] *Tanaka S., Suzuki R., Takahashi T., et al.* Japanese ModernBERT: A Long-Context Encoder for Japanese NLP // arXiv preprint arXiv:2501.17684. – 2025.–P.1–13.DOI:10.48550/arXiv.2501.17684.

[17] *Li Z., Chen X., Wang Y., et al.* Enhancing Retrieval with ModernBERT and ColBERT // arXiv preprint arXiv:2502.04357. – 2025. – P. 1–11. DOI: 10.48550/arXiv.2502.04357.
[18] *Teiletche P., Macé Q., Conti M., et al*. ModernVBERT: Towards Smaller Visual Document Retrievers // arXiv preprint arXiv:2510.01149. – 2025. – P. 1–12. DOI: 10.48550/arXiv.2510.01149.
[19] *Sennrich, R., Haddow, B., & Birch, A.* Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th ACL. 2016. Pages 1715–1725
[20] *Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze.* "Introduction to Information Retrieval." Cambridge University Press (2008): 233-256. doi: 10.1017/CBO9780511809071
[21] *Salemi, Alireza, and Hamed Zamani*. "Evaluating Retrieval Quality in Retrieval-Augmented Generation." arXiv preprint arXiv:2404.13781 (2024): 1-15. doi: 10.48550/arXiv.2404.13781
[22] *Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., & Liu, Y.* Roformer: Enhanced transformer with rotary positional embedding. Neurocomputing. 2024. Vol. 568. Page 127012.
[23] *Beltagy, I., Peters, M. E., & Cohan, A.* Longformer: The Long-Document Transformer. arXiv preprint arXiv:2004.05150. 2020. Pages 1–16
[24] *Hendrycks, D., & Gimpel, K.* Gaussian Error Linear Units (GELU). arXiv preprint arXiv:1606.08415. 2016. Pages 1–8
[25] *Mansurov, B., & Mansurov, A.* UzRoBERTa: A Pretrained Language Model for Uzbek. Journal of Natural Language Processing Challenges. 2021. Pages 45–52.
[26] *Kalyan, K.S., Rajasekharan, A., & Sangeetha, S.* AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing. arXiv preprint arXiv:2108.05542. 2021. Pages 1–35
[27] *Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., & Zettlemoyer, L.* Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th ACL. 2020. Pages 8440–8451
[28] *Reimers, N., & Gurevych, I.* Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the EMNLP-IJCNLP. 2019. Pages 3982–3992
[29] *Xiao, S., Liu, Z., Zhang, J., & Muennighoff, N.* C-Pack: Packaged Resources for General Chinese Embeddings. arXiv preprint arXiv:2309.07597 (BGE-M3 model basis). 2023. Pages 1–15
[30] *Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., ... & Lample, G.* Llama: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971. 2023. Pages 1–19
[31] *Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D*. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. ICLR. 2020. Pages 1–18

Поступила в редакцию 10.01.2026

Citation: *Xujayarov I.Sh., Ochilov M.M., Xolmatov O.A., Jumanov V.I.* (2026). O‘zbek tili uchun kengaytirilgan kontekstli ModernUzBERT semantik embedding modelini ishlab chiqish va samaradorligini baholash. 9(2). – B. 45-53. https://doi.org/10.62132/ijdt.v9i2.375.

DEVELOPMENT AND EVALUATION OF THE MODERNUZBERT SEMANTIC EMBEDDING MODEL WITH EXTENDED CONTEXT FOR THE UZBEK LANGUAGE

Khujayarov I.Sh.[1], Ochilov M.M.[1], + Kholmatov O.A.[1], Jumanov V.I.[2]

[1] Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan
[2] State Institution “Center for the Development of Information and Communication Technologies in Justice Bodies and Institutions”, Tashkent, Uzbekistan

Abstract. This research is dedicated to the development of ModernUzBERT, an advanced embedding model for the Uzbek language based on contemporary Transformer architecture with an extended context window of 8192 tokens. In low-resource and agglutinative languages such as Uzbek, the limited capacity for processing comprehensive context significantly hampers model performance. To address this challenge, a large-scale text corpus in the Uzbek Latin script comprising over 125 million tokens was curated, and an optimized vocabulary of 52,000 sub-word units was developed. The model training process was executed in two primary stages: initially, the model was pre-trained from scratch using the Masked Language Modeling (MLM) objective; subsequently, it was adapted for semantic

search via Supervised Fine-Tuning (SFT) using a dataset of 30,000 question-answer pairs. Through the implementation of Flash Attention 2 and Unpadding technologies, GPU resource utilization efficiency was substantially enhanced. The 8192-token context window enables the analysis of large-scale documents while preserving semantic integrity. Upon conclusion of the study, the developed model and all its components were released for open-access to the scientific community on the Hugging Face platform.

Keywords: NLP, uzbek language, ModernBERT, semantic embedding, information retrieval, long context, flash attention.

РАЗРАБОТКА И ОЦЕНКА ЭФФЕКТИВНОСТИ СЕМАНТИЧЕСКОЙ ЭМБЕДДИНГ-МОДЕЛИ MODERNUZBERT С РАСШИРЕННЫМ КОНТЕКСТОМ ДЛЯ УЗБЕКСКОГО ЯЗЫКА

Хужаяров И.Ш.[1], Очилов М.М.[1], [+] Холматов О.А.[1], Жуманов В.И.[2]

[1] Ташкентский университет информационных технологий имени Мухаммада аль-Хорезми, Ташкент, Узбекистан
[2] Государственное учреждение «Центр развития информационно-коммуникационных технологий в органах и учреждениях правосудия», Ташкент, Узбекистан

Аннотация. Данное исследование посвящено разработке ModernUzBERT - эмбеддинг-модели для узбекского языка, основанной на современной архитектуре Transformer с расширенным контекстным окном (8192 токена). В малоресурсных и агглютинативных языках, таких как узбекский, ограниченные возможности обработки широкого контекста негативно влияют на эффективность моделей. Для решения этой проблемы был сформирован масштабный текстовый корпус на узбекской латинице, содержащий более 125 миллионов словоформ, и создан оптимизированный словарь, состоящий из 52 000 субсимвольных единиц (sub-words). Процесс обучения модели проходил в два этапа: на первом этапе модель обучалась «с нуля» методом маскированного языкового моделирования (Masked Language Modeling, MLM), а на втором - адаптировалась под семантический поиск посредством контролируемого тонкого обучения (Supervised Fine-Tuning, SFT) с использованием датасета из 30 000 вопросно-ответных пар. Благодаря применению технологий Flash Attention 2 и Unpadding удалось значительно повысить эффективность использования графических процессоров (GPU). Контекстное окно в 8192 токена позволяет анализировать объемные тексты, сохраняя их семантическую целостность. По завершении исследования разработанная модель и все её компоненты были выложены в открытый доступ для научного сообщества на платформе Hugging Face.

Ключевые слова: NLP, узбекский язык, ModernBERT, семантические эмбеддинги, информационный поиск, длинный контекст (long context), flash attention.