

TRANSFORMER ARCHITECTURE IN NATURAL LANGUAGE PROCESSING: EVOLUTION FROM BERT TO GPT-4 AND PROSPECTS FOR THE UZBEKISTAN LANGUAGE

Boltayev Ulugbek Zafar ugli

Tashkent University of Information Technologies, Faculty of Software Engineering, 3rd year

boltayevulugbek344@gmail.com

Accepted: 15.02.2024 | Published: 20.04.2024

Annotation: Based on the UzNLP-230M corpus of 230 million words, the UzBERT-base and UzGPT-2 models were trained and evaluated on 8 NLP tasks. UzBERT-base achieved 91.4% F1 in sentiment analysis, 88.7% in named object detection, and 93.2% in text classification — an average of 11.3% better than mBERT. The results provide a clear roadmap for the development of the Uzbek NLP ecosystem.

Keywords: transformer, BERT, GPT, natural language processing, Uzbek language, language model

ISSN: 2815-1003 OKA Jurnal, 2024, 8(2), 15-29

TABIIY TILNI QAYTA ISHLASHDA TRANSFORMERLAR ARXITEKTURASI: BERT DAN GPT-4 GACHA EVOLYUTSIYA VA O'ZBEK TILI UCHUN ISTIQBOLLAR

Boltayev Ulug'bek Zafar o'g'li

Toshkent Axborot Texnologiyalari Universiteti, Dasturiy Injiniroving fakulteti, 3-kurs

boltayevulugbek344@gmail.com

Qabul: 15.02.2024 | Nashr: 20.04.2024

Annotatsiya: 230 million so'zdan iborat UzNLP-230M korpusi asosida UzBERT-base va UzGPT-2 modellari o'qitildi va 8 NLP vazifasida baholandi. UzBERT-base sentiment tahlilida 91.4% F1, nomli ob'ektlarni aniqlashda 88.7%, matn klassifikatsiyasida 93.2% ko'rsatkichlariga erishdi — mBERT dan o'rtacha 11.3% ustun. Natijalar o'zbek NLP ekotizimini rivojlantirish uchun aniq yo'l xaritasini asoslaydi.

Kalit so'zlar: transformer, BERT, GPT, tabiiy tilni qayta ishlash, o'zbek tili, til modeli

I. KIRISH (Introduction)

Vaswani va boshq. (2017) tomonidan taqdim etilgan transformer arxitekturasida NLP sohasida inqilob yasadi. Rekurrent tarmoqlardan (RNN, LSTM) farqli o'laroq, transformerlar parallel hisoblashga imkon berib BERT, GPT-2, T5, GPT-3 va GPT-4 uchun zamin yaratdi. O'zbek tili agglutinativ tuzilishi, morfologik boyligi va ikki alifbosi (lotin/kiril) bilan NLP uchun qo'shimcha qiyinchiliklar taqdim etadi (Tursunov va boshq., 2022).

Tadqiqot gipotezasi: mahalliy O'zbek korpusida noldan o'qitilgan transformer modeli ko'p tilli modellardan barcha NLP vazifalarida statistik jihatdan muhim darajada ustun keladi.

II. METODOLOGIYA (Methods)

2.1. UzNLP-230M korpusi

To'rt manbadan 230M so'z to'plandi: milliy gazetalar arxivi 1999-2023 (87.4M), Vikipediya (34.2M), kutubxona katalogi (62.1M), qonunchilik hujjatlari (46.3M). Preprocessing: HTML tozalash, MinHash dublikat olib tashlash (Jaccard=0.8), unicode normalizatsiya, BPE tokenizatsiya.

2.2. Model arxitekturasini va o'qitish

UzBERT-base: 12 transformer qatlami, 768 yashirin o'lchov, 12 diqqat boshi, 110M parametr, 50,000 tokenli WordPiece lug'at. O'qitish: 8x NVIDIA A100 80GB, 14 sutka, batch 256, lr 1e-4, MLM+NSP maqsadlari. UzGPT-2: 117M parametr, causal language modeling.

2.3. Baholash protokoli

8 NLP vazifasi: sentiment tahlil, NER, matn klassifikatsiya, savol-javob, matn o'xshashligi, so'z ma'nosi ajratish, bog'liqlik tahlili, munosabat ajratish. Taqqoslash modellari: mBERT, XLM-R-base, mGPT.

III. NATIJALAR (Results)

3.1. UzBERT-base natijalari

UzBERT-base barcha 8 vazifada mBERT dan o'rtacha 11.3% yuqori natija ko'rsatdi. Sentiment tahlil: 91.4% F1 (mBERT: 79.2%, XLM-R: 83.7%). NER: 88.7% F1 (mBERT: 74.3%, XLM-R: 80.1%). Matn klassifikatsiya: 93.2% F1 (mBERT: 81.4%, XLM-R: 86.3%). Savol-javob: 76.3% Exact Match (mBERT: 61.2%). Barcha farqlar statistik muhim ($p < 0.001$).

3.2. UzGPT-2 natijalari

Matn generatsiyasida perplexity: UzGPT-2 34.7 vs mGPT 45.1 (23% yaxshilanish). Hallucination darajasi: 18.3%. Inson baholash (5 ball): ravonlik 4.2, faktik to'g'rilik 3.8.

3.3. Domen tahlili

Qonunchilik domenida eng yuqori: 94.1% F1. Ijtimoiy tarmoq matnida eng past: 82.3% F1. Lotin alifbosida 2.3% yuqori natija kiril bilan qiyoslaganda.

IV. MUHOKAMA VA XULOSA (Discussion & Conclusion)

4.1. Muhokama

Natijalar Rust va boshq. (2021) gipotezasini tasdiqlaydi: ko'p tilli modellar o'zbek tilini yomon tokenlashtiradi (o'rtacha 2.3x ko'proq token) va bu samaradorlikni pasaytiradi. UzBERT-base mahalliy WordPiece lug'at orqali ushbu muammoni hal etdi.

4.2. Xulosa

O'zbek tilidagi raqobatbardosh transformer modellari yaratish texnik jihatdan amalga oshirildi. Uchta strategik tavsiya: (1) UzBERT-large va UzLLM-7B uchun milliy hisoblash infratuzilmasi; (2) standart UzGLUE benchmark; (3) lotin-kiril avtomatik normalizatsiya moduli. O'zbekiston 2027 yilga kelib milliy LLM ga ega bo'lish imkoniyatiga ega.

Journal: <https://ifsmrc.org/index.php/aijrm>

Adabiyotlar

1. Brown, T. et al. (2020). Language models are few-shot learners. NeurIPS, 33, 1877-1901.
2. Devlin, J. et al. (2019). BERT: Pre-training of deep bidirectional transformers. NAACL-HLT, 1, 4171-4186.
3. Rust, P. et al. (2021). How good is your tokenizer? ACL-IJCNLP, 4362-4384.
4. Tursunov, A. et al. (2022). Challenges in Uzbek NLP. COLING Workshop, 45-53.
5. Vaswani, A. et al. (2017). Attention is all you need. NeurIPS, 30, 5998-6008.