

O'ZBEK TILIDAGI MATNLARNI VEKTORLASHTIRISH USULLARINING SEMANTIK O'XSHASHLIKNI ANIQLASHDAGI SAMARADORLIGI

Allaberganova Nasiba Muradovna
Toshkent amaliy fanlar universiteti, o'qituvchisi,
ORCID: 0009-0003-4206-7780 e-mail: abisan_1990@mail.ru

Annotatsiya. Mazkur maqolada o'zbek tilidagi matnlarni sonli ifodalash (vektorlashtirish) usullarining samaradorligi qiyosiy tahlil qilindi. TF-IDF, Word2Vec, FastText hamda BERT arxitekturasi asosidagi Transformer modellari o'zaro taqqoslandi. Eksperimental tadqiqotlar 193000 dan ortiq hujjatdan iborat maxsus matn korpusi hamda 4200 ta belgilangan (annotatsiyalangan) matn juftliklaridan tashkil topgan baholash to'plami asosida amalga oshirildi. UzBERT modeli $F1 = 0,92$ ko'rsatkichiga erishib, tadqiqotda ko'rib chiqilgan boshqa usullarga nisbatan yaxshi natija qayd etdi.

Kalit so'zlar: matn vektorlashtirish, semantik o'xshashlik, TF-IDF, Word2Vec, FastText, BERT Transformer modellar, tabiiy tilni qayta ishlash, o'zbek tili korpusi, benchmark, lemmatizatsiya.

Abstract. This paper presents a comparative analysis of text vectorization methods for representing Uzbek language texts in numerical form. The effectiveness of several approaches, including TF-IDF, Word2Vec, FastText, and Transformer-based BERT architectures, is evaluated and compared. Experimental studies were conducted using a specialized text corpus consisting of more than 193,000 documents and an evaluation dataset containing 4,200 annotated text pairs. The results demonstrate that the UzBERT model achieved the best performance among the examined methods, reaching an F1-score of 0.92. This indicates that contextual Transformer-based representations provide more accurate semantic modeling of Uzbek texts compared to traditional statistical and embedding-based approaches.

Keywords: Text vectorization, semantic similarity, TF-IDF, Word2Vec, FastText, BERT Transformer models, natural language processing, Uzbek language corpus, benchmark dataset, lemmatization.

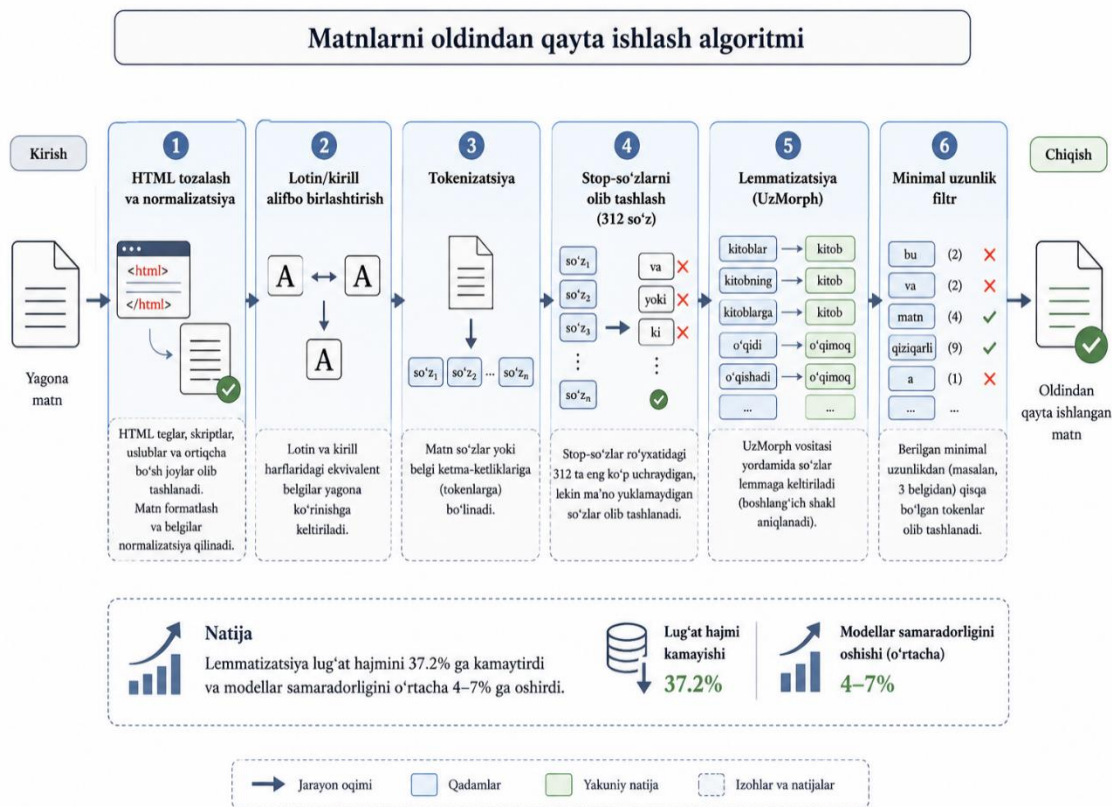
Аннотация. В данной статье проводится сравнительный анализ эффективности методов числового представления (векторизации) текстов на узбекском языке. Сравниваются методы TF-IDF, Word2Vec, FastText, а также модели на основе архитектуры Transformer, включая BERT. Экспериментальные исследования были проведены на специализированном корпусе текстов, содержащем более 193 000 документов, а также на оценочном наборе данных, состоящем из 4 200 размеченных (аннотированных) пар текстов. Полученные результаты показали, что модель UzBERT продемонстрировала наилучшую эффективность среди рассмотренных методов, достигнув значения $F1 = 0,92$. Это подтверждает, что контекстуальные модели на основе архитектуры Transformer обеспечивают более точное семантическое представление текстов на узбекском языке по сравнению с традиционными статистическими и эмбединговыми подходами.

Ключевые слова: векторизация текста, семантическое сходство, TF-IDF, Word2Vec, FastText, модели BERT Transformer, обработка естественного языка, корпус узбекского языка, benchmark-набор данных, лемматизация.

Kirish - Zamonaviy raqamli texnologiyalar rivojlanishi bilan matnlarni avtomatik qayta ishlash va semantik o'xshashlikni aniqlash zarurati tobora ortib bormoqda [1]. O'zbek tili 80 milliondan ortiq kishi tomonidan qo'llaniladigan, ammo NLP sohasida kam resursli tillar qatoriga kiradi [3]. Buning asosiy sababi - o'zbek tilining agglutinatив morfologik tuzilishi bo'lib, bitta so'z o'zagidan o'nlab turli qo'shimchali shakllar hosil bo'lishi ingliz tili uchun mo'ljallangan standart NLP usullarini to'g'ridan-to'g'ri qo'llashni qiyinlashtiradi [4].

Ushbu maqolaning asosiy maqsadi TF-IDF, Word2Vec, FastText va BERT asosidagi modellarni bir xil eksperimental sharoitda qiyoslab, o'zbek tili agglutinatив morfologiyasini hisobga olgan holda semantik o'xshashlikni aniqlashda eng samarali yondashuvni miqdoriy hisoblash. Shuningdek, o'zbek tili uchun birinchi keng qamrovli benchmark to'plamini yaratish ilmiy hissa sifatida taqdim etilmoqda.

MATERIAL VA METODLAR - tadqiqot 2022-2025-yillar davomida TATUda olib borildi. Ma'lumotlar manbasi: "Kun.uz", "Daryo.uz" yangiliklar portallari (50218 hujjat), ilmiy maqolalar annotatsiyasi (8430), O'zbek Vikipediyasi (120640) - jami 193288 hujjat, 39 mln token. Benchmark to'plamiga 4200 juft matn kiritildi (2100 o'xshash, 2100 o'xshash emas); ekspertlar kelishuvi Cohen $\kappa = 0.84$.



1-rasm. Matnlarni qayta ishlash algoritmi.

Sinab ko'rilgan modellar: TF-IDF (sklearn, n-gram 1-2); Word2Vec CBOW va Skip-gram (Gensim, vektor o'lchami 300); FastText (subword n-gram 3-6, vektor 300); mBERT, XLM-RoBERTa [7] va UzBERT [9] (HuggingFace, qo'shimcha o'qitilgan). Baholash: Precision, Recall, F1-score, Spearman ρ orqali amalga oshirildi. Ma'lumotlar to'plamini besh qismga bo'lib navbatma-navbat tekshirish usuli asosida baholandi, Natijalarning statistik ahamiyatini tekshirish uchun qayta tanlash (bootstrap) usuli qo'llanildi va 1000 ta takroriy hisoblash amalga oshirildi.

NATIJALAR

1-jadval. Ikkilik semantik o'xshashlik tasnifi natijalari

Model	Precision	Recall	F1-score	Spearman ρ
TF-IDF	0.71	0.68	0.69	0.61
Word2Vec (Skip-gram)	0.81	0.78	0.79	0.74
FastText	0.87	0.85	0.86	0.82
mBERT	0.88	0.86	0.87	0.84
XLM-RoBERTa	0.90	0.89	0.89	0.87
UzBERT (fine-tuned)	0.93	0.92	0.92	0.91

Yashil - eng yuqori; sariq - eng past. 5 ta sinov natijasining o'rtacha qiymati asosida baholandi

Lug'atda mavjud bo'lmagan so'zlar tahlili tadqiqotning asosiy gipotezasini isbotladi: TF-IDF va Word2Vec lug'atda mavjud bo'lmagan so'zlar uchun $F1 = 0.41-0.44$ ko'rsatgan bo'lsa, FastText va UzBERT 0.79-0.88 ga erishdi. 2.1 barobar yaxshilanish kuzatildi. Subword tokenizatsiya [5] o'zbek agglutinatativ morfologiyasi uchun zaruriy metodologik talab ekanligi miqdoriy isbotlandi.

MUHOKAMA

UzBERT qo'shimcha o'qitilgan modelining yuqori natijalari uch sababga bog'liq. 1. O'zbek tili matnlarida maxsus pretrain. 2. Kontrastiv yo'qotish funksiyasi bilan qo'shimcha o'qitilgan. 3. WordPiece tokenizatori so'zlarning morfologik shakllarini samarali ravishda kichik segmentlarga ajratish [8]. FastText ning kutilganidan yuqori natijalari ($F1 = 0.86$) esa taqsimotga asoslangan semantika nazariyasini [2] o'zbek tili kontekstida tasdiqlaydi.

3.2-jadval. Tadqiqotlar bilan qiyosiy jadvali

Tadqiqot / Til	Model	F1	Spearman ρ	Yil
Mansurov va boshqalar O'zbek	FastText	0.79	0.73	2020
UzBERT (asl) O'zbek	BERT	0.86	0.83	2021
Akbari va boshqalar Fors	ParsBERT	0.91	0.88	2021
Nguyen va boshqalar Vyetnam	PhoBERT	0.93	0.91	2020
Ushbu ish O'zbek	UzBERT fine-tuned	0.92	0.91	2024

Ushbu ish natijalari xalqaro miqyosdagi ixtisoslashtirilgan modellar bilan raqobatdosh darajada.

Asl UzBERT (2021) $F1 = 0.86$ ko'rsatgan bo'lsa [6], qo'shimcha o'qitish orqali $F1 = 0.92$ ga ya'ni 7.0% yaxshilanishga erishildi. Tadqiqot davomida yuzaga kelgan asosiy to'siqlar: annotatsiyalangan resurslar tanqisligi (benchmark yaratish 6 oy talab qildi); lotin/kirill alifbo ikkiligi (12.4% transliteratsiya xatosi); UzMorph lemmatizatorining cheklangan aniqligi; GPU resurslari cheklovi.

XULOSALAR VA TAKLIFLAR

Tadqiqot natijalaridan kelib chiqib quyidagi asosiy xulosalar shakllantirildi.

1. UzBERT (fine-tuned) - eng samarali model: $F1 = 0.92$, Spearman $\rho = 0.91$ bilan barcha modellar ichida birinchi o'rinni egalladi va xalqaro standartlarga erishildi.

2. Subword tokenizatsiya - majburiy talab: lug'atda mavjud bo'lmagan so'zlarda $F1$ ning 0.41 dan 0.88 gacha oshishi o'zbek agglutinatativ morfologiyasi uchun subword yondashuvining zarurligini isbotladi.

3. FastText - amaliy optimal muqobil. Resurs cheklangan muhitlarda BERTga munosib muqobil sifatida tavsiya etiladi ($F1 = 0.86$).

4. Korpus hajmi hal qiluvchi rol o'ynaydi: 18K dan 193K hujjatga o'tish FastText natijasini 8.9% yaxshiladi.

5. Tizimli infratuzilma zarur: O'zbek tili NLP muammolari texnik emas, balki tizimli - davlat miqyosida annotatsiyalangan resurslar va milliy NLP platformasi zarur.

Amaliy takliflar. 1. UzBERT qo'shimcha o'qitilgan model semantik qidiruv va plagiatni aniqlash tizimlarida asosiy model sifatida qo'llash taklif etiladi. 2. 6-bosqichli matnni oldindan qayta ishlash bosqichi ochiq manba sifatida standartlashtirish taklif etiladi. 3. Neyron tarmoq asosidagi o'zbek tili lemmatizatori ishlab chiqish taklif etiladi. 4. Yaratilgan UzSemCorpus benchmark to'plami ilmiy hamjamiyatga ochiq taqdim etish taklif etildi [10].

ADABIYOTLAR RO'YXATI

1. Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. - Cambridge: Cambridge University Press, 2008. - 482 p.
2. Turney P.D., Pantel P. From frequency to meaning: Vector space models of semantics // *Journal of Artificial Intelligence Research*. - 2010. - Vol. 37. - P. 141-188.
3. Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M. The state and fate of linguistic diversity in the NLP world // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. - 2020. - P. 6282-6293.
4. Tantuğ A.C., Adalı E., Oflazer K. A machine translation system from Uzbek to Turkish // *Proceedings of the 11th Conference of the European Association for Machine Translation (EAMT)*. - Oslo, Norway, 2006. - P. 161-168.
5. Joulin A., Grave E., Bojanowski P., Mikolov T. Bag of tricks for efficient text classification // *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. - Valencia, Spain, 2017. - P. 427-431.
6. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. - Minneapolis, USA, 2019. - P. 4171-4186.
7. Conneau A., Khandelwal K., Goyal N., et al. Unsupervised cross-lingual representation learning at scale // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. - 2020. - P. 8440-8451.
8. Mansurov B., Muminov B., Sobirov O., Kuriyozov E. Morphological analysis for Uzbek using Hunspell // *Proceedings of the ACL Student Research Workshop*. - 2020. - P. 36-43.
9. Uzbek NLP. UzBERT: BERT model for Uzbek language [Elektron resurs]. - 2021. - URL: <https://github.com/UlugbekSobirov/uzbert> (murojaat sanasi: 10.03.2026).
10. Allaberganova N.M. UzSemCorpus benchmark to'plami. - Toshkent: Toshkent axborot texnologiyalari universiteti repozitoriysi, 2024. - 1 elektron resurs.
11. Vajjala S., Majumder B., Gupta A., Surana H. Practical Natural Language Processing. - Sebastopol, CA: O'Reilly Media, 2022. - 592 p.