O‘ZBEK TILIDAGI HUJJATLARDAGI XATOLIKLARNI AVTOMATIK ANIQLASHDA MBERT MODELIDAN FOYDALANISH

Axmedova Xolisxon Ilxomovna,
t.f.f.d.PhD
xolisa9029@mail.ru
O‘zbekiston xalqaro islomshunoslik akademiyasi

Mamayusufova Tursunxon Otabek qizi
4-bosqich talabasi
hsn.tursunoy@gmail.com
O‘zbekiston xalqaro islomshunoslik akademiyasi

Annotatsiya. Microsoft Word matn muharririda ingliz tili matnlaridagi imloviy xatoliklarni avtomatik taqdim funksiyasi ishlab chiqilgan. Ammo o‘zbek tilida bunday imkoniyat borligini uchratmadik. Mazkur maqolada o‘zbek tilidagi matnlaridagi imloviy xatolarni aniqlash muammosi tadqiq qilindi. Ushbu masalani yechishda transformer arxitekturaga asoslangan mBERT modelidan foydalanildi. mBERT modeliga imloviy xatolardan holi, tekshirilgan matnlardan iborat datasetda o‘qitildi hamda model hosil qilindi. Taklif etilgan yondashuvga asoslangan model test qilinganda 94% aniqlikka erishildi.

Kalit so‘zlar: *elektron hujjatlar, imlo xatolari, mBert, o‘zbek tili, imlo tekshiruv tizimlari, mashinaviy o‘rganish, tabiiy tilni qayta ishlash.*

Abstract. Microsoft Word text editor features an automatic spell-checking function for English texts. However, we have not encountered such a capability for the Uzbek language. This article investigates the problem of detecting spelling errors in Uzbek language texts. To solve this issue, the mBERT model based on the Transformer architecture was utilized. The mBERT model was trained on a dataset consisting of verified, error-free texts to create the final model. When the proposed approach-based model was tested, it achieved an 94% accuracy rate.

Keywords: *e-documents, spelling errors, mBert, uzbek language, spelling detection systems, Machine Learning, Natural Language Processing (NLP).*

Kirish

Hozirgi kunda axborot texnologiyalarining jadal rivojlanishi natijasida turli tillardagi matnli hujjatlar hajmi yuqori darajada oshib bormoqda. IDC (International Data Corporation) ma’lumotlariga ko‘ra, 2026-yilga kelib jahon bo‘yicha yaratilgan raqamli ma’lumotlar hajmi 221 zettabaytdan oshishi kutilmoqda [1]. Shuningdek, matnlardagi imloviy xatoliklarni aniqlash va tuzatish muammosi axborotning aniq va tushunarli yetkazilishini ta’minlash, NLP tizimlarining samaradorligini oshirish uchun muhim hisoblanadi.

O‘zbek tilida imloviy xatolarni avtomatik aniqlash va tuzatish masalasi bir qator lingvistik hamda texnologik murakkabliklar bilan tavsiflanadi. Jumladan, o‘zbek tilining agglyutinativ xususiyati sababli bitta o‘zakdan yuzlab turli so‘z shakllari hosil bo‘lishi mumkin. Bu esa imloviy xatolarni aniqlash jarayonida so‘zlarning barcha variantlarini qamrab olishni murakkablashtiradi. Bundan tashqari, o‘zbek tilida lotin va kirill yozuvlarining parallel qo‘llanilishi, apostrof belgilarining turlicha yozilishi hamda norasmiy matnlarda transliteratsiya va shevaga xos shakllarning keng uchrashi avtomatik tekshiruv tizimlari aniqligini pasaytiradi. Shu bilan birga, o‘zbek tilida annotatsiyalangan korpuslar va maxsus imloviy xatolar bazasining yetarli emasligi mashinaviy o‘rganish asosidagi modellarni samarali o‘qitishda jiddiy muammo tug‘diradi.

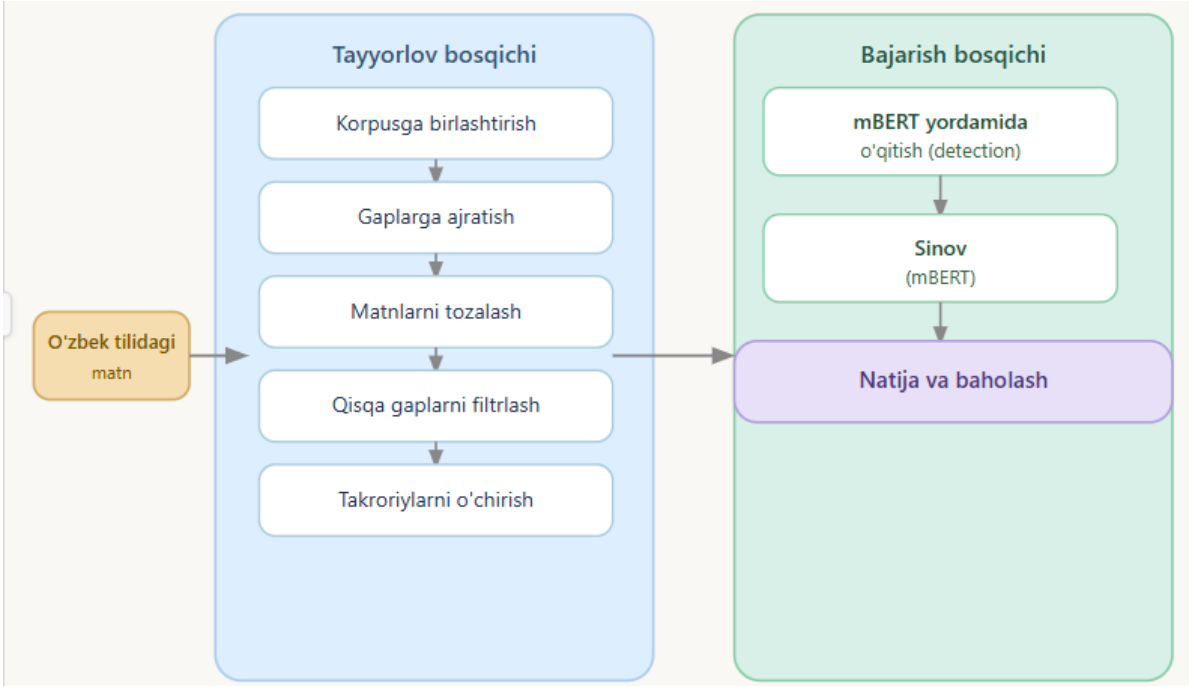
Shu sababli o‘zbek tilidagi matnlarda imloviy xatolarni aniqlash va tuzatish uchun tilning o‘ziga xos xususiyatlarini hisobga oluvchi, zamonaviy mashinaviy o‘rganish algoritmlariga asoslangan yondashuvlarni ishlab chiqish dolzarb ilmiy-amaliy vazifa hisoblanadi. Mazkur tadqiqot aynan shu muammoni hal etishga qaratilgan bo‘lib, o‘zbek tilidagi matnlar uchun samarali va aniq ishlovchi imlo tekshiruv modelini yaratishni maqsad qiladi.

Matnlardagi imloviy xatolarni aniqlash va tuzatish masalasi tabiiy tilni qayta ishlash sohasining muhim yo‘nalishlaridan biri bo‘lib, ushbu muammo bo‘yicha turli yondashuvlar ishlab chiqilgan.

Dastlabki imlo tekshiruv tizimlari lug‘at va grammatik qoidalarga tayangan. Unix muhitida ishlab chiqilgan **Spell** dasturi imloviy xatoliklarni aniqlashga mo‘ljallangan dastlabki va samarali qoidaga asoslangan tizimlardan biri hisoblanadi[2]. Ushbu dastur hujjatni to‘liq kirish ma’lumoti sifatida qabul qilib, texnik matnlarga moslashtirilgan 25.000 so‘zdan iborat lug‘at asosida har bir so‘zni tekshiradi. Lug‘atda mavjud bo‘lmagan so‘zlar imloviy xato sifatida aniqlanadi va foydalanuvchiga ro‘yxat ko‘rinishida taqdim etiladi. Y.Chaabi va F.Ataa Allah Amazigh tili uchun ishlab chiqqan imlo tekshiruv tizimida to‘liq Damerau-Levenshtein va n-gram yondashuvlarini birlashtirib, xatolarni aniqlashda 86.62%–98.74% oralig‘ida F-measure natijalariga erishgan [3]. Shuningdek, Bangla tilidagi tadqiqotlarda edit distance va n-gram modellari kombinatsiyasi 96% o‘rtacha aniqlik ko‘rsatgani qayd etilgan [4]. Singh va hamkorlari tomonidan ishlab chiqilgan **HINDIA** modeli hind tilidagi imloviy xatolarni aniqlash uchun attention-based BiRNN-LSTM arxitekturasidan foydalangan va yuqori aniqlik natijalarini ko‘rsatgan [5]. Assamese tilida olib borilgan tadqiqotda LSTM modeli 92.72%, BiLSTM esa 94.59% aniqlik bilan xatolarni aniqlagan [6]. Tamil tili uchun Uthayamoorthy va boshqalar yaratgan **DDSpell** tizimi 4 million so‘zli korpus asosida bi-gram similarity matching, minimal edit distance va word frequency kombinatsiyasidan foydalangan [7]. Kannada tilida Ramaneedi va boshqalar mT5 modelidan foydalanib matndagi shovqinni 12% kamaytirishga erishgan [8]. Transformer arxitekturasi imlo to‘g‘rilash sohasida yangi bosqichni boshlab berdi. BERT, RoBERTa, T5, NeuSpell, SpellBERT kabi modellar kontekstga bog‘liq xatolarni aniqlashda yuqori samaradorlik ko‘rsatmoqda. Xitoy tili imlo to‘g‘rilash uchun SpellBERT va MDCSpell modellarining natijalari baseline usullardan ustun ekanligi ko‘rsatilgan [9]. Arab tili uchun AraSpell transformer asosidagi samarali model sifatida taklif qilingan [10].

Metodologiya

Taklif etilgan metodologiya uch bosqichdan iborat: ma’lumotlarni oldindan qayta ishlash, bajarish va natijalarni baholash.



1-rasm. Metodologiya bosqichlari

Dataset. Tadqiqot doirasida o‘zbek tilidagi matnlar asosida keng qamrovli dataset shakllantirildi. Ma’lumotlar ikki asosiy manbadan yig‘ildi:

1. Grammatik jihatdan to‘g‘ri tuzilgan badiiy adabiyotlar va elektron hujjatlar;
2. Internet resurslaridan olingan yangilik va maqola matnlari.

Internet ma’lumotlarini yig‘ish jarayoni web scraping texnologiyasi orqali amalga oshirildi. Xatolarni aniqlash vazifasi uchun har bir so‘zga token darajasida ERR (xato) va O (to‘g‘ri) yorliqlari berildi. Yakuniy dataset 65.534 ta gaplardan iborat bo‘lib, modelni o‘qitish uchun yetarli hajmga ega bo‘ldi.

Ma’lumotlarni oldindan qayta ishlash. Yig‘ilgan matnlar dastlab xom ko‘rinishda bo‘lgani sababli, ular bir necha bosqichda qayta ishlanib, model uchun mos formatga keltirildi.

Birinchi bosqichda barcha matnlar yagona korpusga birlashtirildi va ortiqcha belgilar, HTML teglar hamda shovqinli elementlar olib tashlandi. Keyingi bosqichda

matnlar gaplarga ajratildi, juda qisqa va mazmunsiz gaplar datasetdan chiqarib tashlandi hamda takroriy gaplar filtrlandi. Bu dataset sifatini oshirish va modelning umumlashtirish qobiliyatini yaxshilashga xizmat qildi. Keyingi bosqichda imloviy xatolarni aniqlash modelini o‘qitish uchun sun’iy xatolar generatsiya qilindi. Bunda o‘zbek tiliga xos bo‘lgan imloviy xatolar modellashtirildi, jumladan:

- fonetik yaqin harflarning almashishi (o‘ → o, g‘ → g),
- harflarning tushib qolishi,
- harflarning ortiqcha qo‘shilishi,
- qo‘shni harflarning o‘rin almashishi,
- qo‘shimchalar bilan bog‘liq xatolar.

Natijada har bir gap uchun xatoli va to‘g‘ri variantlar shakllantirildi hamda token darajasida belgilash amalga oshirildi. Har bir token “O” (to‘g‘ri) yoki “ERR” (xatoli) yorlig‘i bilan belgilandi.

Yakuniy bosqichda ma’lumotlar modelga mos formatga keltirildi. Tokenizatsiya jarayonida matnlar alohida so‘zlarga ajratildi va mBERT modelining tokenizeri yordamida raqamli ko‘rinishga o‘tkazildi. Natijada tozalangan, strukturaviy jihatdan tartiblangan va belgilangan dataset hosil qilindi. Ushbu dataset imloviy xatolarni aniqlash modelini o‘qitish va baholash uchun asosiy ma’lumot manbasi sifatida xizmat qildi.

mBert asosida imloviy xatolarni aniqlash modeli

Mazkur tadqiqotda imloviy xatolarni aniqlash uchun quyidagicha yondashuv qo‘llanildi:

Xatolarni aniqlash vazifasi token klassifikatsiya muammosi sifatida qaraldi. Ushbu bosqichda bert-base-multilingual-cased modeli asosida qurilgan mBERT modeli qo‘llanildi. Ushbu model transformer arxitekturasi asosida qurilgan bo‘lib, matndagi so‘zlarning kontekstual bog‘liqligini chuqur o‘rganish imkonini beradi. mBERT modeli ko‘p tilli katta hajmdagi korpuslarda oldindan o‘qitilgan bo‘lib, u

turli tillarda, jumladan o‘zbek tilidagi matnlar bilan ishlashda samarali natijalar beradi.

Model arxitekturasi kiruvchi matnni tokenlarga ajratish va har bir token uchun kontekstga bog‘liq vektor hosil qilishga asoslanadi. Ushbu vektorlar orqali model har bir tokenning gap ichidagi semantik va sintaktik rolini aniqlaydi. mBERT modelining asosiy ustunligi shundaki, u har bir so‘zni alohida emas, balki butun gap kontekstida tahlil qiladi. Bu esa imloviy xatolarni aniqlashda muhim ahamiyatga ega.

Modelni amalga oshirish jarayonida token darajasida klassifikatsiya yondashuvi qo‘llanildi. Ya’ni, har bir token ikki sinfdan biriga tegishli deb belgilandi:

- **O** – to‘g‘ri yozilgan token;
- **ERR** – imloviy xatoga ega token.

Tokenizatsiya jarayonida mBERT tokenizeridan foydalanildi.

Modelni o‘qitish va testlash. Dataset uch qismga ajratildi: o‘quv (training dataset), tekshiruv (validation dataset) va test to‘plamlari. Bu modelni ortiqcha moslashuvdan (overfitting) himoya qilish va natijalarni obyektiv baholash imkonini berdi.

mBERT modeli token tasniflash vazifasi uchun bir nechta epoch davomida o‘qitildi. Har bir gap tokenlarga ajratilib, tegishli teglar bilan birga modelga uzatildi. Model kontekstga asoslangan xususiyatlarni o‘rganish orqali xato va to‘g‘ri tokenlarni ajratishni o‘zlashtirdi.

Test qilish bosqichida esa modelga yangi, ilgari modelga taqdim etilmagan matnlar berildi va uning aniqligi baholandi. Har bir token uchun model ehtimollik qiymatini hisoblab, oldindan belgilangan threshold asosida yakuniy sinfga ajratdi. Agar tokenning xato bo‘lish ehtimoli ma’lum chegaradan yuqori bo‘lsa, u **ERR** sinfiga kiritildi.

```
test_sentence = "Men bugun maktapga bordim va togri javob yozdim"

results = detect_errors_in_sentence(test_sentence)

for r in results:
    print(r)
```

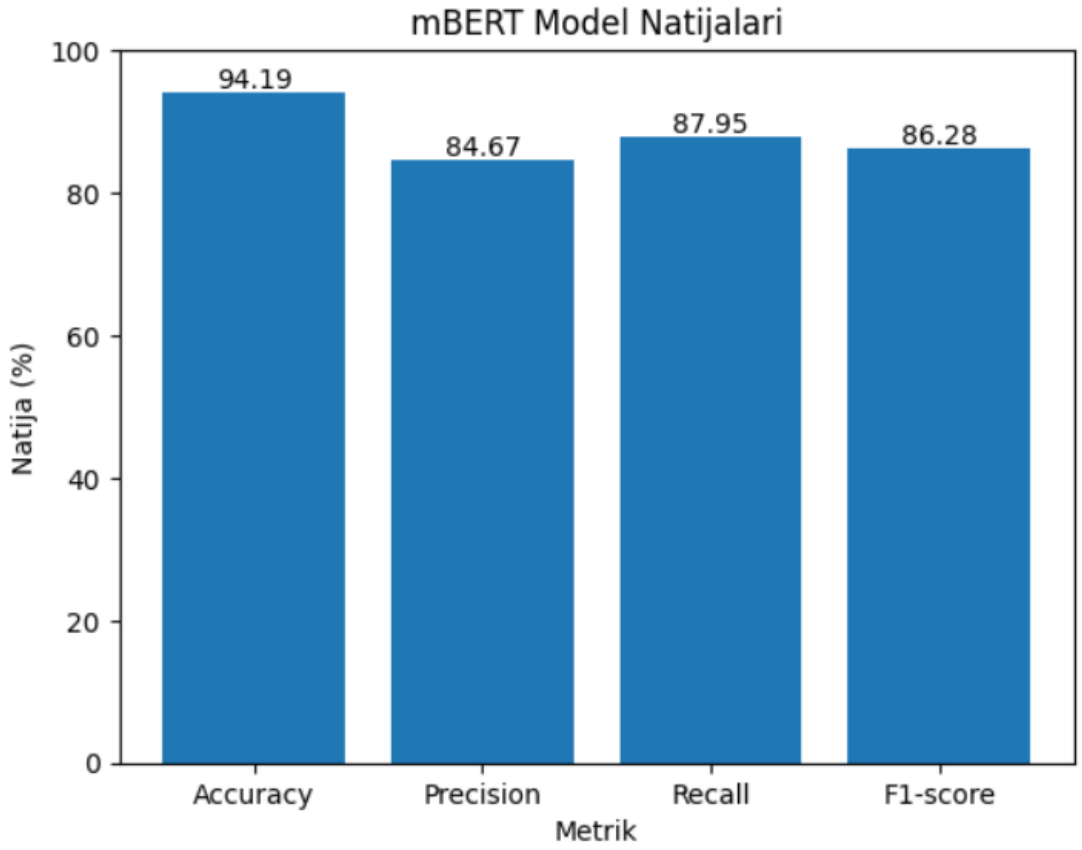
```
{'token': 'Men', 'err_probability': 0.0, 'label': 'O'}
{'token': 'bugun', 'err_probability': 0.0023, 'label': 'O'}
{'token': 'maktapga', 'err_probability': 0.9647, 'label': 'ERR'}
{'token': 'bordim', 'err_probability': 0.1379, 'label': 'O'}
{'token': 'va', 'err_probability': 0.0001, 'label': 'O'}
{'token': 'togri', 'err_probability': 0.9998, 'label': 'ERR'}
{'token': 'javob', 'err_probability': 0.0001, 'label': 'O'}
{'token': 'yozdim', 'err_probability': 0.0007, 'label': 'O'}
```

2-rasm. Modelning sinov natijasi

Modelning kontekstni hisobga olgan holda ishlashi uning an’anaviy usullarga nisbatan ustunligini ko‘rsatadi hamda imloviy xatolarni aniqlash masalasida chuqur o‘rganish yondashuvlarining samaradorligini tasdiqlaydi.

Tadqiqot natijalari va tahlili

Tajriba ishlari davomida mBERT asosidagi token darajasida ishlovchi model o‘qitildi va alohida test to‘plamida sinovdan o‘tkazildi. Eksperimental natijalarga ko‘ra, model test to‘plamida quyidagi natijalarga erishdi:



3-rasm. mBert model natijalari

Grafikdan ko‘rinib turibdiki, model eng yuqori natijani umumiy aniqlik bo‘yicha ko‘rsatgan. Shuningdek, ehtimollik chegarasini moslashtirish orqali model samaradorligini yanada oshirish imkoniyati mavjudligi aniqlandi. Optimal threshold qiymati tanlanganda modelning F1-score ko‘rsatkichi **86.98%** gacha oshdi.

Olingan natijalar shuni ko‘rsatadiki, modelning yuqori aniqlik ko‘rsatkichlari uning amaliy qo‘llanilish imkoniyatlarini tasdiqlaydi.

Xulosa

Mazkur tadqiqotda o‘zbek tilidagi matnlarda imloviy xatolarni aniqlashga mo‘ljallangan avtomatik imlo tekshiruv modeli ishlab chiqildi. Tadqiqot doirasida badiiy adabiyotlar hamda internet manbalaridan yig‘ilgan matnlar asosida keng qamrovli dataset shakllantirildi va u real hamda sun’iy xatolar bilan boyitildi.

Umuman olganda, taklif etilgan yondashuv o‘zbek tilida imloviy xatolarni aniqlash muammosini samarali hal qilish imkonini beradi. Shu bilan birga, kelgusida

real foydalanuvchi ma’lumotlari asosida datasetni kengaytirish va modelni yanada takomillashtirish orqali uning aniqligini oshirish mumkin.

Foydalanilgan adabiyotlar ro‘yxati

1. Bagchi P., Arafin M., Akther A., Alam K.M. Bangla Spelling Error Detection and Correction Using N-Gram Model. In: International Conference on Machine Intelligence and Emerging Technologies. Springer; 2022. Pp. 468–82.

2. Big Data Statistics 2026 (Growth, Trends, Market Size), December 29, 2025.

3. Chaabi Y., Ataa Allah F. Amazigh spell checker using Damerau-Levenshtein algorithm and N-gram. Journal of King Saud University Computer and Information Sciences.

4. Liu S. et al. CRASpell: a contextual typo robust approach to improve Chinese spelling correction. Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, Dublin, Ireland.

5. Phukan R. Automated Spelling Error Detection in Assamese Texts using Deep Learning Approaches.

6. Ramaneedi S., Pati PB. Kannada Textual Error Correction Using T5 Model. In: 2023 IEEE 8th International Conference for Convergence in Technology (I2CT). IEEE; 2023. p. 1–5.

7. Salhab M., Abu-Khzam F. AraSpell: a deep learning approach for Arabic spelling correction. IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), 2023 p.1-16. https://doi.org/10.48550/arXiv.2405.06981

8. Singh S., Singh S. HINDIA: a deep-learning-based model for spell-checking of Hindi language. Neural Computing and Applications. 2021;33:3825–40.

9. Uthayamoorthy K., Kanthasamy K., Senthaalan T., Sarveswaran K., Dias G. Ddspell: a data-driven spell checker and suggestion generator for the Tamil

language. In: 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE; 2019. p. 1–6.