# O'ZBEK TILIDA MATNLARNING SENTIMENTAL TAHLILI: ROBERTA VA T5 MODELLARINI SOLISHTIRISH

*Uskanova Dilafruz[1], Davronov Rifqat[2]*

*[1]Aniq va Ijtimoiy fanlar universiteti*
*[2]Romanovskiy nomidagi matematika institute*
*E-mail: imomovnalolaa@gmail.com*

K E Y W O R D S

sentimental tahlil, o'zbek tili, transformator modellari, RoBERTa, T5, matnni klassifikatsiya qilish, tabiiy tillarni qayta ishlash (NLP)

A B S T R A C T

Mazmuniy tahlil (sentiment analysis) — tabiiy tillarni qayta ishlash (NLP) sohasining eng faol rivojlanayotgan yo'nalishlaridan biri bo'lib, u matnlarda ifodalangan his-tuyg'u va munosabatlarni avtomatik aniqlashga xizmat qiladi. Hozirgi kunda ingliz va boshqa yirik tillar uchun yuqori natijaga erishgan transformator arxitekturali modellar o'zbek tili kabi resurslar cheklangan tillar uchun hali etarli darajada o'rganilmagan. Ushbu maqolada o'zbek tilidagi 29 194 ta yozuvdan iborat annotatsiyalangan dataset asosida ikki xil chuqur o'rganish modeli — RoBERTa (BERT asosidagi ketma-ket klassifikator) va UzT5 (T5 asosidagi generativ model) — solishtirilib tekshiriladi. Tajriba natijalari shuni ko'rsatadiki, RoBERTa arxitekturasi asosidagi model 76,5% aniqlik (accuracy) va 75,7% F1-ko'rsatkichi bilan UzT5 modelidan (52,7% aniqlik, 52% F1) sezilarli darajada ustun keladi. Olingan natijalar o'zbek tili uchun diskriminativ klassifikatsiya yondashuvining generativ yondashuvga nisbatan samaradorligini tasdiqlaydi.

## 1. Kirish

Raqamli kommunikatsiyaning jadal rivojlanishi natijasida ijtimoiy tarmoqlar, onlayn do'konlar va ilovalar sharhlari tarzida ulkan hajmdagi foydalanuvchi fikrlari to'planmoqda. Ushbu ma'lumotlardan samarali foydalanish uchun mazmuniy tahlil — matnlardagi his-tuyg'ularni (ijobiy, salbiy yoki neytral) avtomatik aniqlash — katta ahamiyat kasb etadi. Mazmuniy tahlil texnologiyalari biznes razvedkasi, mahsulot sifatini baholash, siyosiy tahlil va jamiyat kayfiyatini kuzatishda keng qo'llanilmoqda [1].

Hozirgi kunda ingliz tili uchun BERT [2], RoBERTa [3], GPT [4] kabi kuchli transformator modellari mavjud bo'lib, ular ko'pgina NLP vazifalarida insonning natijalariga yaqinlashib qoldi. Biroq o'zbek tili kabi resurslar cheklangan (low-resource) tillar uchun bunday modellar va annotatsiyalangan ma'lumot to'plamlari soni hamon juda kam. Buning sabablari orasida: o'zbek tilining morfologik murakkabligi (aglyutinativ tuzilish), yozuv tizimlari xilma-xilligi (lotin, kirill), va sifatli annotatsiyalangan korpuslarning yo'qligi bor [5].

Ushbu tadqiqot quyidagi asosiy maqsadlarni ko'zlaydi:

1. O'zbek tilidagi ilovalar sharhlari asosida sentimental tahlil uchun ixtisoslashgan dataset yaratish;
2. BERT asosidagi diskriminativ model (RoBERTa) va T5 asosidagi generativ model (UzT5) xatti-harakatini o'zbek tilida solishtirish;
3. Ikki xil modellashtirish paradigmasini (fine-tuning for classification va seq2seq generation)

(cc) BY 

taqqoslash orqali o'zbek tili uchun eng samarali yondashuvni aniqlash.

## 2. Adabiyotlar Sharhi

### 2.1 Sentimental Tahlilning Rivojlanishi

Sentimental tahlil sohasidagi ilk ishlar leksika va qoidalarga asoslangan usullarni qo'llagan (SentiWordNet, VADER). Chuqur o'rganishning (deep learning) rivojlanishi bilan LSTM va CNN asosidagi modellar yuqori natijalar ko'rsata boshladi. Ammo transformator arxitekturasining [6] kashf etilishi va BERT [2] modeli taqdim etilishi bilan soha tubdan o'zgardi. BERT modeli oldindan o'qitish (pre-training) va nozik sozlash (fine-tuning) paradigmasi orqali ko'pgina NLP vazifalarida inqilob yasadi.

RoBERTa (Robustly Optimized BERT Pre-training Approach) [3] modeli BERT'ning takomillashtirilgan versiyasi bo'lib, kattaroq ma'lumot to'plami va optimal gipegiperparametrlar asosida o'qitilgan. T5 (Text-To-Text Transfer Transformer) [7] esa barcha NLP vazifalarini matndan-matnga aylantiruvchi yagona freymvork sifatida taklif qilingan va "so'z + ketma-ket chiqish" yondashuvi bilan klassifikatsiya, tarjima, xulosa chiqarish kabi vazifalarni bajaradi.

### 2.2 O'zbek Tili uchun NLP Tadqiqotlari

O'zbek tili uchun NLP tadqiqotlari nisbatan yosh bo'lib, so'nggi yillarda sezilarli ilgarilanishlar ro'y berdi. UzBERT [8] — o'zbek tili uchun maxsus o'qitilgan BERT modeli — morfologik tahlil va nomlangan birikmalarni aniqlash vazifalarida yuqori natija ko'rsatdi. UzT5 modeli esa o'zbek tilida generativ vazifalar uchun, shu jumladan matn klassifikatsiyasi uchun sinab ko'rilgan. Shu bilan birga, o'zbek tilida sentimental tahlilga bag'ishlangan tadqiqotlar soni cheklangan bo'lib, bu borada standartlashtirilgan benchmark dataset ham deyarli yo'q.

### 2.3 Mavjud Muammolar

O'zbek tilidagi mazmuniy tahlilning asosiy muammolari: - **Morfologik murakkablik:** O'zbek tili aglyutinativ bo'lib, bitta so'z ko'plab affiks qabul qilishi mumkin; - **Ko'p qirralilik:** Bir xil so'z turli kontekstda qarama-qarshi ma'no anglatishi mumkin; - **Ma'lumot tanqisligi:** Tasdiqlangan, katta hajmli annotatsiyalangan dataset yo'qligi; - **Yozuv ikki tizimliligi:** Lotin va kirill yozuvlarining parallel ishlatilishi.

## 3. Ma'lumotlar To'plami

### 3.1 Datasetning Tavsifi

Ushbu tadqiqotda fix_uz_dataset_latin30k.csv fayli asosida yaratilgan dataset ishlatildi. Dataset o'zbek tilida yozilgan ilovalar (mobil dasturlar) sharhlari to'plamidan iborat bo'lib, lotin yozuvida taqdim etilgan.

**Jadval 1.**

**Dataset statistikasi**

| Ko'rsatkich | Qiymat |
|---|---|
| Jami yozuvlar soni | 29 194 |
| Sentimental sinflar | 3 (Ijobiy, Neytral, Salbiy) |
| Dublikatlar | 1 222 |
| Tozalangandan keyingi yozuvlar | 27 972 |
| Ustunlar | idx, Score (sinflar), Text |

### 3.2 Sinflar Taqsimoti

Dataset ma'lumotlarini tahlil qilganda, sinflar orasidagi taqsimot quyidagicha ekanligi aniqlandi:

**Jadval 2.**

**Sinflar taqsimoti (tozalashdan oldin)**

| Sinf | Namunalar soni | Ulush (%) |
|---|---|---|
| Salbiy (Negative) | 13 043 | 44,7% |
| Ijobiy (Positive) | 10 682 | 36,6% |
| Neytral (Neutral) | 5 469 | 18,7% |
| **Jami** | **29 194** | **100%** |

Dataset ma'lum darajada nomutanosib bo'lib, salbiy sinfda namunalar soni ko'proq. Bu holat real hayotdagi ilovalar sharhlariga xos bo'lib,

(cc) BY 

foydalanuvchilar ko'pincha muammo yuz berganida sharh yozadi.

**Namuna yozuvlar:** - *Ijobiy:* "Uni yaxshi ko'raman" — "Juda ham ajoyib. Bugungi kunga qadar men 36 ta…" - *Salbiy:* "Karaoke ishlamayapti. O'zgarish kerak." — "Juda tez donduruyor Juda tez-tez muzlashadi." - *Neytral:* "Bepul bepul qo'shiq"

### 3.3 Ma'lumotlarni Tozalash va Taqsimlash

Ma'lumotlarni tayyorlashda quyidagi bosqichlar bajarildi:

1. Dublikatlarni olib tashlash: drop_duplicates(subset='Text') orqali 1 222 ta takroriy yozuv o'chirildi (29 194 → 27 972);
2. Sinflarni kodlash: Matn sinflar (Positive, Neutral, Negative) raqamli ko'rinishga o'tkazildi: {Negative: 0, Neutral: 1, Positive: 2};
3. Ma'lumotlarni bo'lish:
4. RoBERTa uchun: Trening — 17 527 (59%), Validatsiya — 5 536 (19%), Test — 6 131 (21%), jami 29 194 ta yozuv ustida (takrorlanmagan);
5. UzT5 uchun: Trening — 23 063 (79%), Test — 6 131 (21%).

## 4. Metodologiya

### 4.1 RoBERTa Modeli (Diskriminativ Klassifikatsiya)

RoBERTa — BERT'ning takomillashtirilgan varianti bo'lib, ushbu tadqiqotda o'zbek tili uchun moslashtirilgan RoBERTa-Base modeli ishlatildi. Model AutoModelForSequenceClassification sinfi orqali yuklab olinib, uch sinfli (3-label) klassifikatsiya uchun sozlandi. Model chiqish qatlami softmax faollashtirish funksiyasi orqali har bir sinf uchun ehtimollik qiymatini qaytaradi.

Tokenizatsiya uchun AutoTokenizer ishlatildi. Har bir matn 512 token uzunligiga qisqartirildi yoki to'ldirildi (truncation=True, padding='max_length'). Bu transformator modellarida standart yondashuv hisoblanadi.

**Jadval 3.**

**RoBERTa fine-tuning sozlamalari**

| Parametr | Qiymat |
|---|---|
| O'qish tezligi (learning rate) | $2\times10^{-5}$ |
| Batch hajmi (trening/validatsiya) | 16 / 16 |
| Maksimal epochlar soni | 50 |
| Og'irlik pasayishi (weight decay) | 0,01 |
| Baholash strategiyasi | Har bir epoch |
| Erta to'xtash (early stopping) | Sabr: 2 epoch |
| Eng yaxshi model yuklash | Ha |

Erta to'xtash mexanizmi (EarlyStoppingCallback) validatsiya yo'qotishining ketma-ket 2 ta epochda yaxshilanmasligi holatida trening jarayonini to'xtatadi. Bu ortiqcha moslashishning (overfitting) oldini oladi.

Model quyidagi mezonlar bo'yicha baholandi: - Aniqlik (Accuracy): To'g'ri tasniflangan namunalar ulushi; - F1-ko'rsatkichi (Macro): Har bir sinf uchun F1 o'rtacha qiymati (sinf nomutanosibligi uchun muvofiq); - Aniqlik darajasi (Precision, Macro): Prognoz qilingan pozitivlar orasida haqiqiy pozitivlar ulushi; - To'liqlik (Recall, Macro): Haqiqiy pozitivlarning aniqlanish ulushi.

### 4.2 UzT5 Modeli (Generativ Klassifikatsiya)

UzT5 — o'zbek tili uchun maxsus oldindan o'qitilgan T5 modeli (19GB o'zbek matni asosida). Tadqiqotda T5 sentimental tahlil uchun noan'anaviy — generativ — tarzda ishlatildi: model kirishdagi matnni qabul qilib, chiqishda sinf raqamini (0, 1 yoki 2) matn ko'rinishida generatsiya qiladi. Bu yondashuv AutoModelForSeq2SeqLM sinfi orqali amalga oshirildi.

- Kirish matnlar (text) tokenizatsiya qilindi va uzunligi 1347 tokenga qadar bo'lishi mumkinligi aniqlandi;
- Chiqish yorliqlari (label: "0", "1", "2") tokenizatsiya qilindi — maksimal uzunlik 3 token;
- Padding tokenlari yo'qotish funksiyasida -100 bilan almashtirildi.

(cc) BY

**Jadval 4.**

**UzT5 fine-tuning sozlamalari**

| Parametr | Qiymat |
|---|---|
| O'qish tezligi (learning rate) | $3\times10^{-4}$ |
| Batch hajmi (trening/test) | 4 / 4 |
| Maksimal epochlar soni | 50 |
| Prognozni generatsiya qilish | Ha |
| Eng yaxshi model yuklash | Yo'q |
| Saqlash strategiyasi | Har bir epoch (max 2) |

## 5. Tajriba Natijalari

### 5.1 RoBERTa Natijalari

Model 3 ta epoch davomida o'qitilgandan so'ng erta to'xtash mexanizmi ishga tushdi va eng yaxshi model yuklandi.

**Jadval 5.**

**RoBERTa baholash natijalari (test to'plami)**

| Mezon | Qiymat |
|---|---|
| Aniqlik (Accuracy) | **76,55%** |
| F1-ko'rsatkichi (Macro) | **75,68%** |
| Aniqlik darajasi (Precision, Macro) | **76,08%** |
| To'liqlik (Recall, Macro) | **78,97%** |
| Baholash yo'qotishi (Eval Loss) | 0,5222 |

Trening jarayoni tahlili:

- 1-epoch: Validatsiya yo'qotishi — 0,5222; Accuracy — 76,55%; F1 — 75,68%
- 2-epoch: Validatsiya yo'qotishi — 0,5319; Accuracy — 76,45%; F1 — 73,93%

Validatsiya yo'qotishi 2-epochda oshgani uchun erta to'xtash ishga tushdi va 1-epochning eng yaxshi modeli saqlab qolindi. Bu natijalarga ko'ra, model nisbatan tez konvergentsiyaga erishgan.

### 5.2 UzT5 Natijalari

UzT5 modeli generativ yondashuv bilan sentimental tahlilni bajarish uchun o'qitildi. 6 131 ta test namunasi ustida qo'l bilan baholash amalga oshirildi.

**Jadval 6.**

**UzT5 baholash natijalari (test to'plami)**

| Sinf | Aniqlik darajasi | To'liqlik | F1 | Namunalar |
|---|---|---|---|---|
| Salbiy (0) | 0,54 | 0,67 | 0,59 | 2 765 |
| Neytral (1) | 0,45 | 0,33 | 0,38 | 1 163 |
| Ijobiy (2) | 0,55 | 0,45 | 0,50 | 2 203 |
| **Aniqlik** | — | — | — | **52,7%** |
| **Weighted avg** | 0,52 | 0,53 | 0,52 | 6 131 |

### 5.3 Modellarni Solishtirish

**Jadval 7.**

**RoBERTa va UzT5 modellarini umumiy solishtirish**

| Mezon | RoBERTa | UzT5 | Farq |
|---|---|---|---|
| Aniqlik (Accuracy) | **76,55%** | 52,68% | +23,87 % |
| F1-ko'rsatkichi | **75,68%** | 52,0% | +23,68 % |
| Aniqlik darajasi (Macro) | **76,08%** | ~50,0% | — |
| To'liqlik (Macro) | **78,97%** | ~48,3% | — |
| Parametrlar soni | ~125M | ~220M | — |
| Moslashuv yondashuvi | Klassifikatsiya | Generativ | — |

RoBERTa modeli barcha asosiy ko'rsatkichlarda UzT5 modelidan 23-24 foiz ustun kelgan.

## 6. Muhokama

### 6.1 RoBERTa ustunligining sabablari

RoBERTa modelining UzT5 dan keskin ravishda yaxshi natija ko'rsatishi quyidagi omillar bilan izohlanadi:

1. Arxitektura mosliq: AutoModelForSequenceClassification klassifikatsiya uchun maxsus optimallashtirilgan bo'lib, model chiqishi to'g'ridan-to'g'ri sinf ehtimolliklarini o'rganadi. Bu diskriminativ yondashuv klassifikatsiya vazifalariga to'g'ridan-to'g'ri mos keladi.
2. Generativ yondashuvning kamchiliklari: UzT5 modeli asosan generativ vazifalar (tarjima, xulosa, javob berish) uchun mo'ljallangan. Uni klassifikatsiya uchun ishlatishda model ba'zan raqam o'rniga boshqa belgilar yoki bo'sh qatorlar generatsiya qilishi mumkin, bu esa noto'g'ri tasniflashga olib keladi.
3. Tokenizatsiya effektivligi: RoBERTa uchun maksimal uzunlik 512 token bilan cheklangan va barcha namunalar bir xil tarzda paddinglangan. UzT5 uchun esa maksimal kirish uzunligi 1 347 tokenga yetgan, bu esa hisoblash murakkabligini va xato ehtimolini oshiradi.
4. Batch hajmi: UzT5 uchun batch hajmi 4 ta bilan cheklangan (GPU xotirasi tufayli), bu esa trening barqarorligini pasaytiradi. RoBERTa uchun esa 16 ta batch bilan o'qitildi.

### 6.2 Neytral sinfning murakkabligi

Har ikkala model ham neytral sinfni eng kam aniqlik bilan tasniflagan. Bu holat o'zbek tilidagi neytral matnlarning ko'pincha ham ijobiy, ham salbiy elementlarni o'z ichiga olishidan kelib chiqadi. Bundan tashqari, neytral sinf namunalari soni boshqa sinflarga nisbatan ancha kam (5 469 ta = 18,7%), bu esa model uchun yetarlicha o'rgatuvchi misol bo'lmaydi.

### 6.3 O'zbek tili xususiyatlarining ta'siri

O'zbek tilining aglyutinativ tabiatidan kelib chiqib, so'zlar juda ko'p shakl qabul qilishi mumkin. Misol uchun, "yaxshi" so'zi "yaxshiroq", "yaxshilash", "yaxshilanmoqda" kabi ko'pgina hosilalarni berishi mumkin. RoBERTa tokenizatori subword darajasida ishlash orqali bu muammoni qisman bartaraf etadi. O'zbek tiliga moslashtirilgan tokenizatorning mavjudligi model samaradorligini oshiruvchi asosiy omillardan biri bo'ldi.

### 6.4 Cheklovlar

Ushbu tadqiqotning asosiy cheklovlari: - Dataset faqat mobil ilovalar sharhlari doirasida cheklangan; - O'qitish jarayonida batch hajmi GPU xotirasi tufayli cheklangan; - UzT5 modeli uchun to'liq o'qitish (50 epoch) tugatilmagan.

## 7. Xulosa

Ushbu tadqiqotda o'zbek tilidagi 29 194 ta annotatsiyalangan matndan iborat dataset asosida ikki modelning sentimental tahlil samaradorligi solishtirilib o'rganildi. Asosiy xulosalar:

1. RoBERTa (diskriminativ) yondashuv o'zbek tili sentimental tahlili uchun 76,55% aniqlik va 75,68% F1-ko'rsatkichi bilan yetakchi natija ko'rsatdi;
2. UzT5 (generativ) yondashuv klassifikatsiya vazifasi uchun 52,68% aniqlik bilan uncha samarali emas — generativ modellar klassifikatsiyaga bevosita moslantirilmagani sababli pastroq natija berdi;
3. Dataset tahlili shuni ko'rsatdiki, neytral sinf har ikkala model uchun ham eng qiyin kategoriya bo'ldi — bu sinf oz namamunali va kontekstual jihatdan belirsiz;
4. Sinflar nomutanosibligi (Negative: 44,7%, Positive: 36,6%, Neutral: 18,7%) modellarning neytral va ijobiy sinflarni aniqlashiga salbiy ta'sir ko'rsatishi mumkin.

### Kelajakdagi Ishlar

- Kattaroq va muvozanatlashtirilgan o'zbek tili dataset yaratish;
- Turli domenlardagi (yangiliklar, ijtimoiy tarmoqlar, tibbiyot) matnlarda modelni sinash;
- Aspect-Based Sentiment Analysis (ABSA) — muayyan jihatlar bo'yicha sentimental tahlilni amalga oshirish;
- Sinflar nomutanosibligini bartaraf etish uchun data augmentation va sinflarga og'irlik berish usullarini qo'llash;
- LLM (Large Language Model) asosidagi zero-shot va few-shot yondashuvlarini solishtirish.

(cc) BY 

**Adabiyotlar**

1. Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT 2019*, 4171-4186.
3. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.
4. Brown, T., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
5. Abdurakhmonova, N., & Matlatipov, G. (2022). O'zbek tilini kompyuter asosida qayta ishlash: muammolar va istiqbollar. *O'zbekiston Milliy universiteti xabarlari*.
6. 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*, 30.
7. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 21(140), 1-67.
8. Mansurov, B., Mansurov, A., & Ikhtiyor, R. (2021). UzBERT: O'zbek tili uchun oldindan o'qitilgan til modeli. *arXiv preprint arXiv:2110.12693*.

(cc) BY 