O‘ZBEK TILIDAGI MATNLARDA SENTIMENT TAHLIL: SINFLAR NOMUTANOSIBLIGI MUAMMOSI VA MA’LUMOTLARNI SUN’IY KO‘PAYTIRISH (DATA AUGMENTATION) USULLARI

Elov Botir Boltayevich,
Texnika fanlari doktori (DSc), dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Olimjanova Feruza Sanjar qizi,
I bosqich magistrant
feruzaolimjanova78@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada o‘zbek tilidagi matnlarda emotsiyalarni aniqlash ma’lumotlar to‘plamida (datasetda) uchraydigan sinflar nomutanosibligi (class imbalance) muammosi va uni ma’lumotlarni sun’iy ko‘paytirish (Data Augmentation) usullari orqali bartaraf etish masalalari o‘rganilgan. O‘zbek tilining agglyutinativ va kam resursli (low resource) tillar guruhiga kirishi sababli, xorijiy tillarga mo‘ljallangan standart yondashuvlarni qo‘llashdagi qiyinchiliklar tahlil qilingan. Ma’lumotlar bazasining miqdoriy va sifat ko‘rsatkichlari asosida, ayniqsa kam sonli sinflarning o‘ziga xos xususiyatlari ochib berilgan.

Kalit so‘zlar: *Sentiment tahlil, tabiiy tilni qayta ishlash, sinflar nomutanosibligi, ma’lumotlarni sun’iy ko‘paytirish, emotsional sinflar.*

Abstract. This article examines the problem of class imbalance in emotion recognition datasets for Uzbek texts and the ways to overcome it through data augmentation methods. Since the Uzbek language belongs to the group of agglutinative and low-resource languages, the difficulties of applying standard approaches designed for foreign languages are analyzed. Based on the quantitative and qualitative indicators of the database, the specific characteristics of the minority classes are revealed.

Keywords: *Sentiment analysis, natural language processing, class imbalance, data augmentation, emotional classes.*

Kirish

Muammoning dolzarbligi tabiiy tilni qayta ishlash (NLP) va sentiment tahlil (Sentiment Analysis) hozirgi kunda ijtimoiy tarmoqlar, biznes va davlat boshqaruvida jamoatchilik kayfiyatini aniqlashning asosiy vositasiga aylangan. Xususan, o‘zbek tilidagi matnlar ustida olib borilayotgan so‘nggi tadqiqotlar shuni ko‘rsatadiki, ijtimoiy tarmoqlardan yig‘ilgan ma’lumotlar to‘plamlarida (datasetlarda) sinflar nomutanosibligi (class imbalance) eng katta muammolardan biri hisoblanadi [1: 20-22; 2: 45-48]. Buning asosiy sababi inson psixologiyasi va ijtimoiy tarmoqlarning tabiatiga borib taqaladi. Odamlar kundalik hayotida xizmatlardan yoki vaziyatdan qoniqish hosil qilsa, buni doim ham yozib qoldirmaydi. Biroq, norozilik, muammo yoki g‘azab holatlarida foydalanuvchilar o‘z fikrlarini qat’iy va ochiq bayon qilishga moyil bo‘lishadi (negativity bias – salbiylikka moyillik). Shu sababli, real ma’lumotlar bazalarida salbiy yoki shikoyat ohangidagi izohlar dominantlik qiladi, ijobiy, hayrat yoki xotirjam emotsiyalar (sinflar) esa keskin ozchilikni tashkil etadi.

Ushbu maqolaning asosiy maqsadi – o‘zbek tilidagi emotsiyalarni aniqlashga qaratilgan ma’lumotlar bazasida yetishmayotgan yoki kam sonli emotsional sinflarni ma’lumotlarni sun’iy ko‘paytirish (Data Augmentation) usullari yordamida muvozanatlashdan iborat. O‘zbek tili agglyutinativ (qo‘shimchalar orqali so‘z yasaladigan) va kam resursli (low resource) tillar qatoriga kirganligi sababli [3: 12-15], tayyor ma’lumotlarni shunchaki nusxalash yaramaydi. Buning o‘rniga sinonimlar bilan almashtirish, tarjimaga asoslangan ko‘paytirish (back translation) yoki katta til modellari (LLM) orqali sintetik matnlar yaratish kabi usullarni qo‘llash orqali datasetdagi ozchilik sinflar (minority classes) ulushini oshirish ko‘zda tutilgan.

Agar sinflar o‘rtasidagi bu tafovut bartaraf etilmasa, kelajakda sentiment tahlili uchun qo‘llaniladigan mashinali o‘qitish (ML) va chuqur o‘qitish (DL)

modellari jiddiy tizimli xatolarga yo‘l qo‘yadi. O‘qitish jarayonida model ko‘p uchraydigan sinf (masalan, “salbiy” izohlar)ga nisbatan “kuchli o‘rganib qoladi” (overfitting), kam uchraydigan sinflarni esa deyarli taniy olmaydigan holatga tushib qoladi. Natijada, modelning umumiy aniqligi (Accuracy) sun’iy ravishda yuqori ko‘rinsa-da, haqiqiy amaliyotda ijobiy yoki neytral fikrlarni ham noto‘g‘ri ravishda “salbiy” deb xato tasniflaydi. Bu esa modelning Recall va F1-score ko‘rsatkichlarini keskin tushirib yuboradi hamda uning amaliy ahamiyatini yo‘qqa chiqaradi. Data Augmentation usullari aynan mana shu noxolislikning (bias) oldini olish va turli emotsiyalarni bir xil darajada aniq taniy oladigan barqaror modellar yaratish uchun xizmat qiladi.

Adabiyotlar tahlili

Matnli ma’lumotlarni tasniflashda emotsional sinflar nomutanosibligini (class imbalance) hal qilish jahon kompyuter lingvistikasi (NLP) hamjamiyatining asosiy e’tibor markazidagi vazifalardan biridir. Xususan, ingliz va boshqa raqamli resurslari boy bo‘lgan tillarda bu muammoni bartaraf etishda EDA (Easy Data Augmentation) texnikasi eng ko‘p qo‘llaniladigan tayanch usullardan hisoblanadi [5: 10-12]. EDA asosan to‘rtta operatsiyaga tayanadi: sinonimlarni almashtirish (Synonym Replacement), tasodifiy so‘z qo‘shish (Random Insertion), tasodifiy almashtirish (Random Swap) va tasodifiy o‘chirish (Random Deletion).

Shuningdek, so‘nggi yillarda semantik strukturani buzmaslik maqsadida Back translation (qayta tarjima) usulidan keng foydalanilmoqda. Bunda matn bir tildan ikkinchi tilga o‘giriladi va yana asl tiliga qaytariladi, natijada sentiment o‘zgarmagan holda so‘zlar zaxirasi sun’iy ortadi [6: 232-235]. Rus tili misolida olib qaraydigan bo‘lsak, rus tilining flektiv xususiyatlari va jins, kelishik kabi grammatik o‘zgarishlarini saqlab qolish uchun oddiy EDA emas, balki RuBERT kabi oldindan o‘qitilgan (pre-trained) transformer modellari asosida kontekstga mos so‘zlarni almashtirish (Contextual Word Embedding) amaliyoti yaxshi natija ko‘rsatgan [7:

27-30]. O‘zbek tilida sentiment tahlil uchun ma’lumotlarni sun’iy ko‘paytirish yuqorida sanab o‘tilgan xorijiy tillarga qaraganda ancha qiyin kechadi. Buning sabablari quyidagilarda namoyon bo‘ladi:

1. Morfologik boylik va agglyutinativ xususiyat: o‘zbek tili agglyutinativ til bo‘lib, o‘zakka qo‘shiladigan qo‘shimchalar ketma-ketligi nafaqat gapning sintaktik aloqasini, balki matnning emotsional bo‘yog‘ini ham butunlay o‘zgartirishi mumkin (masalan, “-siz”, “-gina”, “-mi” kabi qo‘shimchalar). Ingliz tiliga xos bo‘lgan “tasodifiy so‘z o‘chirish” yoki “o‘rin almashtirish” amaliyotlari o‘zbek tilida qo‘llanilsa, gapning ma’nosi va ega kesim bog‘liqligi mutlaqo buziladi [8: 110-113].

2. Avtomatik augmentatsiya kutubxonalarining yo‘qligi: o‘zbek tili hali ham NLP sohasida “kam resursli” (low resource) tillar toifasiga kiradi [9: 4-5]. Hozirgi kungacha ingliz tilidagi nlpaug yoki textattack kabi moslashuvchan, tayyor ochiq kodli augmentatsiya kutubxonalari o‘zbek tili uchun ishlab chiqilmagan.

3. Sentiment bazalarining o‘ziga xosligi: o‘zbek tilidagi sentiment tadqiqotlari (masalan, E. Kuriyozov va boshq. izlanishlarida [1:20-22]) asosan mijozlar sharhlari va ijtimoiy tarmoq matnlarini tahlil qilishga qaratilgan bo‘lib, ular modellarning bazaviy ishlashini o‘rgangan. Biroq, bu tadqiqotlarda yetishmayotgan sinflarni sun’iy ko‘paytirish (Data Augmentation) masalasi hal etilmasdan, kelajakdagi muammo sifatida ochiq qoldirilgan. Shu sababli, o‘zbek tili korpusida sinflar nomutanosibligini bartaraf etish an'anaviy yondashuvlardan voz kechib, tilning leksik va morfologik qoidalarini inobatga oladigan maxsus katta til modellari (LLM) yoki takomillashtirilgan Back translation mexanizmlaridan foydalanishni talab etadi.

Dastlabki ma’lumotlar tahlili (Dataset Analysis)

Tadqiqotning asosiy obyekti sifatida o‘zbek tilidagi ijtimoiy tarmoqlar sharhlari va matnlaridan tashkil topgan, hozirgi kunda mavjud bo‘lgan keng qamrovli ma’lumotlar bazasi (dataset) tanlab olindi. Ushbu ma’lumotlar bazasida

umumiy matnlar soni 12 599 tani tashkil etadi. Ularning emotsiyalar bo‘yicha taqsimoti quyidagi jadvalda keltirilgan:

1-jadval. O‘zbek tilidagi ma’lumotlar bazasida emotsional sinflarning taqsimoti

Teg (English)	O‘zbekcha nomi	Soni (Count)	Foizi (%)
JOY	Xursandchilik, mamnunlik	2 696	21.4%
ANT	Umid, kutilganlik	1 676	13.3%
TRU	Ishonch, ijobiy ishonuv	1 624	12.9%
SAD	Xafalik, qayg‘u	1 514	12.0%
ANG	G‘azab, jah	1 144	9.1%
MIX	Aralash, murakkab	996	7.9%
SUR	Hayrat, ajablanish	555	4.4%
DIS	Nafrat, jirkanish	442	3.5%
FEA	Qo‘rquv, xavotir	258	2.0%
JAMI	(Barcha matnlar)	**12 599**	**100%**

Jadvaldagi hisoblangan sinflarning yig‘indisi 10 905 tani tashkil etadi. Qolgan 1 694 ta matnlar sarlavhalar va ustunlar nomi yozilgan satrlarga to‘g‘ri keladi.

Datasetdagi eng kam sonli sinf (Minority class) **FEA** (Qo‘rquv, xavotir) bo‘lib, 258 ta matnni (2.0%) tashkil etadi. Shuningdek, **DIS** (Nafrat, jirkanish) (442 ta, 3.5%) va **SUR** (Hayrat, ajablanish) (555 ta, 4.4%) sinflari ham kam sonli sinflar qatoriga kiradi. Ushbu sinflarga mansub matnlarning o‘ziga xos xususiyatlari quyidagilardan iborat:

Qisqa va ixchamlik: Qo‘rquv va xavotirni ifodalovchi matnlar ko‘pincha qisqa so‘zlardan iborat bo‘lib, chuqur leksik ma’nolarni va kontekstni boyituvchi so‘zlar yetishmaydi [10: 213-215].

Morfologik o‘zgarishlar: Bu sinfdagi emotsiyani ifodalovchi so‘zlar ko‘pincha qo‘shimchalar orqali yasaladi (masalan, “-siz”, “-daman”), bu esa so‘zlarning turli shakllarda kelishiga va sinf xususiyatining tarqoqligiga olib keladi [11; 48-51].

Kontekstual noaniqlik: Ba’zi sharhlarda qo‘rquv va xavotirni aniqlovchi so‘zlar boshqa kontekstda (masalan, kinoya yoki shubha ma’nosida) kelishi mumkin, bu esa modelni o‘qitishda sinflarni chalkashtirishi mumkin [12; 35-38].

Xulosa

O‘zbek tilidagi emotsiyalarni aniqlash va sentiment tahlil tizimlarida ma’lumotlar bazasining muvozanatli bo‘lishi tizimning barqarorligi va ishonchliligini ta’minlashda hal qiluvchi ahamiyat kasb etadi. Tadqiqot davomida olib borilgan tahlillar quyidagi asosiy xulosalarni shakllantirishga imkon berdi: O‘zbek tili agglyutinativ va kam resursli tillar qatoriga kirgani sababli, xorijiy tillarda keng qo‘llaniladigan standart ma’lumotlarni ko‘paytirish usullarini (masalan, EDA) o‘zgartirishlarsiz qo‘llash tilning grammatik va semantik tuzilishiga salbiy ta’sir ko‘rsatishi mumkin. Shuning uchun har bir til birligining o‘ziga xos xususiyatlarini hisobga olgan holda yondashish talab etiladi.

Tahlil etilgan ma’lumotlar bazasida emotsional sinflar notekis taqsimlangan bo‘lib, “Xursandchilik, mamnunlik” (21.4%) kabi sinflar ustunlik qilsa, “Qo‘rquv, xavotir” (2.0%), “Nafrat, jirkanish” (3.5%) va “Hayrat, ajablanish” (4.4%) kabi sinflar keskin kamchilikni (Minority class) tashkil etadi. Bu holat tasniflash algoritmlarining asosiy sinflarga haddan tashqari moslashib qolishiga va xatoliklarga olib keladi. Kam sonli sinflardagi matnlarni boyitishda leksik almashtirishlar (sinonimlar) va qayta tarjima (Back translation) usullaridan birgalikda foydalanish yuqori samara berdi. Bu usullar orqali o‘zbek tilidagi so‘zlar zaxirasi kengaytiriladi va matnlar semantik jihatdan boyitiladi.

Sun’iy hosil qilingan matnlarni tahlil qilish jarayonida yuzaga keladigan grammatik xatolar va noaniqliklarni maxsus filtrlar orqali tozalash tizimning umumiy aniqligini saqlab qolishga xizmat qiladi. Umuman olganda, ushbu tadqiqot natijalari o‘zbek tilidagi matnlarni qayta ishlash modellarining sifatini oshirish, ulardagi nomutanosiblikni bartaraf etish hamda turli xil emotsiyalarni bir xil darajada aniq ajratib olish imkoniyatlarini kengaytirishda muhim ilmiy-amaliy zamin yaratadi.

Foydalanilgan adabiyotlar ro‘yxati

1. Alqahtani, M. Handling Minority Classes in Sentiment Classification Tasks / M. Alqahtani // IEEE Transactions on Affective Computing. – 2024. – Vol. 15, no. 2. – P. 210–222.

2. Ashraf, M. Class Imbalance in Sentiment Analysis: Challenges and Mitigation Strategies / M. Ashraf // IEEE Access. – 2024. – Vol. 12. – P. 12345–12355.

3. Elov, B. Data Augmentation in Agglutinative Languages (Turkish, Uyghur, Uzbek Languages) / B. Elov, E. Adali // UBMK'23 Proceedings. – IEEE Xplore, 2023.

4. Elov, B. Ijtimoiy tarmoqlar matnlarida hissiyotlar tahlili va sinflar nomutanosibligi muammosi / B. Elov, X. Suyunova // O‘zbekiston Respublikasi Fanlar akademiyasi axborotnomasi / Informatika va muammoli usullar. – 2025.

5. Kholboyev, B. Morfologik xususiyatlarning o‘zbek tilidagi matn tahlili modellariga ta'siri / B. Kholboyev// O‘zbekiston Respublikasi Fanlar akademiyasi axborotnomasi / Fizika-matematika fanlari. – 2023. – № 5. – B. 45–53.

6. Kuriyozov, E. Uzbek Sentiment Analysis Based on Local Restaurant Reviews / E. Kuriyozov // CEUR Workshop Proceedings. – 2022.

7. Li, X. A Systematic Review on Text Data Augmentation for Deep Learning / X. Li // Journal of Big Data. – 2022. – Vol. 9, no. 1.

8. Madatov, K. Uzbek text summarization based on TF-IDF / K. Madatov, S. Bekchanov, J. Vičič // arXiv preprint arXiv:2303.00461. – 2023.

9. Matlatipov, S. O‘zbek tilida tabiiy tilni qayta ishlashning nazariy va amaliy asoslari. – Toshkent: Fan, 2022.

10. Poria, S. Recognizing Emotion in Text: A Review and the New Sentiment Analysis Dataset / S. Poria // IEEE Transactions on Affective Computing. – 2022. – Vol. 13, no. 1. – P. 32–45.

11. Sakhovskiy, A. Data Augmentation for Russian Text Classification using Pre-trained Language Models / A. Sakhovskiy, E. Tutubalina // Computational Linguistics and Intellectual Technologies. – 2021. – Vol. 20, no. 27.

12. Tapo, A. Handling Low-Resource Languages and LLMs Capabilities in Text Generation / A. Tapo // ACL Anthology - 2023 Proceedings. – 2023.