O‘ZBEK TILIDA SENTIMENT TAHLILI UCHUN AVTOMATLASHTIRILGAN LEKSIK RESURS YARATISH

Abdumalikova Gulshoda Shuhrat qizi
magistrant
abdumalikovagulshoda2403@gmail.com
Toshkent davlat sharqshunoslik universiteti

Annotatsiya. Ushbu maqolada o‘zbek tili agglyutinativ tabiatini hisobga olgan holda sentiment tahlili uchun avtomatlashtirilgan lug‘at yaratish metodologiyasi ishlab chiqilgan. Tadqiqotda 10 000 ta gapdan iborat uch qatlamli (axborot, ijtimoiy va rasmiy matnlar) korpusini shakllantirish, mBERT va GPT modellarini qo‘llash hamda Word2Vec va PMI statistik metodlari orqali leksik birliklarning hissiy koeffitsiyentini [-1; +1] oralig‘ida hisoblash bosqichlari ko‘rsatib o‘tilgan. Taklif etilayotgan avtomatlashtirilgan tizim an’anaviy qo‘lda shakllantiriladigan lug‘at yaratish jarayoniga nisbatan vaqt unumdorligini 1000 baravardan ortiqroqqa oshirishi va tahlil aniqligini 85% gacha yetkazishi asoslab berilgan.

Kalit so‘zlar. *o‘zbek tili, sentiment tahlili, NLP, avtomatlashtirilgan leksik resurslar, PMI metodi, Word2Vec, korpus lingvistikasi, lemmatizatsiya.*

Abstract. This article develops a methodology for creating an automated dictionary for sentiment analysis, accounting for the agglutinative nature of the Uzbek language. The research outlines the formation of a three-layered corpus consisting of 10,000 sentences (news, social, and official texts), the application of mBERT and GPT models, and the calculation of emotional coefficients for lexical units in the range of [-1; +1] using Word2Vec and PMI statistical methods. It is demonstrated that the proposed automated system increases time efficiency by more than 1000 times compared to manual dictionary creation and improves analysis accuracy to 85%.

Keywords: *Uzbek language, sentiment analysis, NLP, automated lexical resources, PMI method, Word2Vec, corpus linguistics, lemmatization.*

Kirish. Hozirgi kunda tabiiy tillarni qayta ishlash (NLP) sohasida matnlar tonalligini aniqlash, ya’ni sentiment tahlili eng jadal rivojlanayotgan yo‘nalishlardan biri hisoblanadi. Axborot oqimining globallashuvi sharoitida raqamli matnlardagi hissiy bo‘yoqni avtomatik aniqlash ijtimoiy fikrni monitoring qilish va foydalanuvchi tajribasini o‘rganishning asosiy vositasiga aylandi. Ushbu tadqiqotning ahamiyati shundaki, sifatli sentiment lug‘atiga ega bo‘lish nafaqat akademik lingvistik tahlillar uchun, balki ilmiy tadqiqotlar, biznes tahlili, siyosiy texnologiyalar va davlat boshqaruvida fuqarolar murojaatlari bilan ishlash tizimlarini avtomatlashtirish uchun o‘ta muhimdir. O‘zbek tili uchun bunday resurslarning yaratilishi milliy muhitdagi sun’iy intellekt tizimlarining intellektual salohiyatini oshirishga xizmat qiladi. Sentiment tahlili sohasida ingliz tili misolida Bing Liu[2;] va Christopher Manning[3] kabi olimlar tomonidan fundamental metodologiyalar ishlab chiqilgan. Turkiy tillar, xususan, o‘zbek tili kontekstida Baxtiyor Mengliyev[4], Shahlo Hamroyeva[6] kabi olimlar va boshqa tadqiqotchilar tomonidan morfologik analizatorlar va korpus lingvistikasi bo‘yicha sezilarli natijalarga erishilgan bo‘lsa-da, avtomatlashtirilgan leksik resurslarni (sentiment lug‘atlarini) generatsiya qilish masalasi hali ham o‘z yechimini kutayotgan sohalardan biri bo‘lib qolmoqda. Mavjud tadqiqotlarning aksariyati lug‘atlarni qo‘lda shakllantirishga yoki boshqa tillardagi resurslarni to‘g‘ridan-to‘g‘ri tarjima qilishga asoslangan. Biroq, bu yondashuv o‘zbek tilining agglyutinativ tabiati va ko‘p ma’noli so‘zlarning hissiy yukini to‘liq qamrab ololmaydi. Mazkur maqola ushbu bo‘shliqni to‘ldirish maqsadida, tayanch so‘zlar va distributiv semantika metodlaridan foydalangan holda, inson omilisiz kengayib boruvchi lug‘at yaratish metodologiyasini taklif etadi. Bu tadqiqotchilar uchun o‘zbek tilidagi resursi kam sohalarni boyitishning yangi uslubiy asosi bo‘lib xizmat qiladi.

Asosiy qism. Taklif etilayotgan metodologiya jarayonda sentiment lug‘atlarni avtomatik tarzda shakllantirish quyidagi bosqichlarni qamrab oladi: ma’lumotlarni to‘plash, matnlarga dastlabki ishlov berish, sentiment darajalanishini hisoblash.

Ma’lumotlarni to‘plash. Ushbu jarayon ma’lumotlarni yig‘ishdan boshlanib, natijalarni validatsiya qilishgacha bo‘lgan texnik amallarni qamrab oladi.

Tadqiqot doirasida o‘zbek tilining turli leksik qatlamlarini qamrab olish maqsadida jami 10 000 dan ortiq gaplardan iborat bo‘lgan korpus shakllantiriladi. Ma’lumotlar bazasini boyitishda quyidagi uchta asosiy axborot manbasidan foydalanish rejalashtirilgan:

1. Axborot matnlar: “Kun.uz” va “Daryo.uz” kabi ommabop yangiliklar portallarining ijtimoiy tarmoqlardagi (Telegram, Facebook) kanallaridan foydalanuvchi sharhlari yig‘ib olinadi. Bu qatlam ijtimoiy faol leksika va norasmiy nutq uslubini tahlil qilish imkonini beradi.
2. Ijtimoiy matnlar: YouTube kabi elektron platformalaridagi videolar ostida qoldirilgan sharhlar to‘planadi. Ushbu manba ijobiy va salbiy qutblanishni aniq ko‘rsatuvchi terminlarga boyligi bilan ajralib turadi.
3. Rasmiy axborot manbalari: Davlat idoralari va axborot agentliklarining rasmiy bayonotlari(Lex.uz) orqali korpusdagi neytral va akademik leksika muvozanati ta’minlanadi.

Ma’lumotlarni yig‘ish jarayoni Python dasturlash tilining BeautifulSoup va Selenium kutubxonalari yordamida, veb-skraping usuli orqali amalga oshiriladi. To‘plangan gaplar soni bo‘yicha taqsimot quyidagicha belgilanishi rejalashtirilgan:

- Axborot matnlar: 4 000 dan oshiq gap;
- Ijtimoiy matnlar: 4 000 dan oshiq gap;
- Rasmiy axborot manbalari: 3 000 dan oshiq gap.

Matnlarga dastlabki ishlov berish. O‘zbek tili agglyutinativ til bo‘lgani sababli, to‘plangan 10 000 dan ziyod gapni to‘g‘ridan-to‘g‘ri lug‘atga kiritish

imkonsiz. Shu sababli, to‘plangan gaplarni lug‘at tarkibiga kiritishdan avval ularni lingvistik jihatdan saralash va hissiy qiymatlarini belgilash jarayoni amalga oshiriladi. Ushbu bo‘lim quyidagi mantiqiy bosqichlardan iborat bo‘ladi:

1) Tozalash va segmentatsiya. Dastlab, matnlar turli texnik shovqinlardan (URL, emoji, maxsus belgilar) tozalanadi va tahlil qulayligi uchun alohida gaplar ko‘rinishida ajratiladi.

2) Sun’iy intellekt yordamida teglash: Yig‘ilgan 10 000 dan ortiq gapning hissiy qutblari (ijobiy, salbiy yoki neytral) sun’iy intellekt modellari (masalan, mBERT yoki GPT asosidagi o‘zbek tili uchun sozlangan modellar) yordamida avtomatik ravishda annotatsiyalanadi.

3) Yordamchi so‘zlar(stop-words)ni filtrlash. Hissiy ma’no yukiga ega bo‘lmagan yordamchi so‘zlar (va, bilan, uchun, hamda va b.) korpusdan avtomatik ravishda chiqarib olinadi.

4) Hissiy yukka ega leksikani ajratish. Markirovka qilingan (ijobiy, salbiy va neytral) gaplar ichidan statistik metodlar (masalan, TF-IDF yoki Chi-square test) yordamida eng ko‘p ahamiyatga ega bo‘lgan kalit so‘zlar avtomatik ajratib olinadi.

5) Lemmatizatsiya va lug‘atni shakllantirish: ajratib olingan kalit so‘zlar metodologiyasi asosida lemma shaki ya’ni o‘zakka keltiriladi va ularning hissiy koeffitsiyentlari asosida yakuniy sentiment lug‘ati generatsiya qilinadi.

10 000 ta gapdan iborat korpusdagi so‘zlar orasidagi semantik yaqinlikni aniqlash uchun Word2Vec yoki FastText modellari o‘qitiladi. Ushbu modellar har bir so‘zni ko‘p o‘lchovli vektor fazosida ifodalaydi va belgilarni avtomatik hisoblash uchun eng avvalgi jarayonlarni o‘z ichiga oladi.

Tayanch so‘zlarga vektor fazosida eng yaqin bo‘lgan yangi so‘zlar Kosinus o‘xshashligi metriksasi yordamida aniqlanadi:

$$similarity(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$$

Bunda tayanch so‘zga semantik jihatdan eng yaqin bo‘lgan yangi leksik birliklar (masalan, *“zo‘r”* so‘zi orqali *“alo”* so‘zi) avtomatik ravishda lug‘atga qo‘shiladi.

Sentiment darajalanishini hisoblash. Lug‘atni shakllantirishning yakuniy bosqichida har bir ajratib olingan leksik birlikka uning hissiy kuchini ifodalovchi sonli qiymat biriktirilishi rejalashtirilgan. Ushbu jarayon quyidagi mantiqiy va matematik asoslarga ko‘ra amalga oshiriladi:

1) Matnli ma’lumotlar tarkibidagi hissiy bo‘yoqqa ega so‘zlar turli darajadagi ta’sir kuchiga ega. Masalan, *“yomon”*, *“dahshatli”* va *“rasvo”* leksemalarining barchasi salbiy qutbga mansub bo‘lsa-da, ularning hissiy intensivligi turlicha. Shu sababli, har bir lemmaga [-1; +1] oralig‘idagi tarozi pallasida o‘z vaznini belgilash ko‘zda tutilgan.

2) Statistik hisoblash algoritmi leksik birliklarning hissiy vaznini aniqlashda PMI metodologiyasidan foydalanish rejalashtirilmoqda. Algoritm korpusdagi ijobiy va salbiy deb teglangan gaplar ichida ma’lum bir so‘zning uchrash ehtimolligini quyidagi formula orqali hisoblab chiqadi:

$$PMI(word, sentiment) = \log_2 \frac{P(word, sentiment)}{P(word)P(sentiment)}$$

Ushbu formula yordamida har bir so‘zning hissiy moyilligi quyidagicha aniqlanadi:

– Agar so‘z asosan ijobiy kontekstli gaplarda (masalan, *“sifatli”* so‘zi 95% holatda ijobiy gaplarda) uchrasa, unga yuqori ijobiy koeffitsiyent (+0.9) biriktiriladi.

– Agar so‘z ko‘p hollarda salbiy gaplar tarkibida kelsa, uning qiymati salbiy qutbga (-0.8) yo‘naltiriladi.

– Ikkala kontekstda ham bir xil chastotada uchraydigan neytral so‘zlar (masalan, *“bugun”*) 0.00 qiymatiga ega bo‘ladi.

3) Kutilayotgan natijalar jadvali: algoritmik hisob-kitoblar yakunida shakllanadigan raqamli lug‘atning taxminiy ko‘rinishi quyidagi jadvalda tasvirlangan (1-jadval):

1-jadval

Leksik birlik (Lemma)	Sentiment koeffitsiyenti	Tavsifi
A’lo	+0.95	O‘ta ijobiy
Yaxshi	+0.50	Ijobiy
Chidasa bo‘ladi	+0.20	Kuchsiz ijobiy
Doska / Bugun	0.00	Neytral birlik
Nuqsonli	-0.60	Salbiy
Rasvo	-0.98	O‘ta salbiy

4) Shakllantirilgan ushbu lug‘at yangi kiruvchi gaplarning umumiy tonalligini aniqlashda matematik asos bo‘lib xizmat qiladi. Masalan, *“Telefon yaxshi (+0.50), lekin ekrani rasvo (-0.98)”* shaklidagi gapning umumiy sentimenti lug‘atdagi qiymatlarning yig‘indisi (yig‘indi=-0.48) orqali avtomatik ravishda salbiy deb tasniflanadi.

Leksik resursning hajmi va qamrovi. Ushbu tadqiqotda kutilayotgan ko‘rsatkichlar o‘zbek va boshqa turkiy tillar bo‘yicha o‘tkazilgan avvalgi NLP tadqiqotlari natijalariga asoslangan. Turkiy tillar korpusi ustidagi tahlillari shuni ko‘rsatadiki, agglyutinativ tillarda 10 000 ta gapdan iborat tanlanma o‘rtacha 15 000-20 000 ta noyob so‘z shakllarini hosil qiladi. Taklif etilayotgan lemmatizatsiya bosqichidan so‘ng, ushbu birliklar o‘zaklarga keltirilganda, hissiy yukka ega bo‘lgan lemmalar soni taxminan 3 000 tadan 5 000 tagacha bo‘lishi bashorat qilinmoqda. Bu ko‘rsatkich B.Liu [2:155] tomonidan ingliz tili uchun tavsiya etilgan o‘rtacha sentiment lug‘ati hajmiga mutanosibdir. Sayfullayeva va Abdurahmonova[1:6] tomonidan 2025-yilda o‘tkazilgan tadqiqotda o‘zbek tili uchun qo‘llanilgan sentiment analiz modellarida 80-82% atrofida aniqlik qayd etilgan. Taklif etilayotgan modelga ko‘ra Word2Vec distributiv semantikasi va PMI statistik vaznlashtirish algoritmining kombinatsiyasi qo‘llanilishi hisobiga, aniqlik ko‘rsatkichini 85% dan yuqori bo‘lishini kutmoqdamiz. Bu distributiv semantika

metodining so‘zlar orasidagi kontekstual bog‘liqlikni aniqlashdagi yuqori samaradorligi bilan izohlanadi.

Ushbu lug‘at o‘z ichiga ijtimoiy-siyosiy munosabatni ifodalovchi terminlar(2-jadval)ni oladi:

2-jadval

№	Termin	Sentiment qiymati	Izoh
1	Adolat	+0.9	Kuchli ijobiy ijtimoiy qadriyat
2	Tengsizlik	−0.8	Salbiy ijtimoiy holat
3	Hamkorlik	+0.7	Ijobiy munosabat va jarayon
4	Nazorat	−0.2	Ko‘pincha cheklov ma’nosida
5	Islohot	+0.3	Kontekstga bog‘liq, yengil ijobiy
6	Byurokratiya	−0.6	Salbiy baholanadi
7	Fuqarolik faolligi	+0.8	Demokratik qadriyat
8	Mojaro	−0.7	Salbiy ijtimoiy hodisa
9	Barqarorlik	+0.6	Ijobiy holat
10	Propaganda	−0.9	Kuchli salbiy diskurs

Xulosa. Tajribali lingvist mutaxassis tomonidan 5 000 ta leksik birlikning (lemma) hissiy qutbi va intensivligini (scoring) aniqlash jarayonini ko‘rib chiqamiz. Har bir so‘zni kontekstual tahlil qilish va unga [-1; +1] oralig‘ida qiymat biriktirish uchun o‘rtacha $T_{manual} = 1$ daqiqa vaqt sarflanishi hisobga olinsa, jami vaqt sarfi T_{total} quyidagicha bo‘ladi:

$$T_{total} = 5000 \times 1\ min = 5000\ min \approx 83.3\ soat$$

Bu esa bir mutaxassisning tanaffuslarsiz qariyb 10 ish kunini tashkil etadi.

Avtomatlashtirilgan algoritm unumdorligi odatiy qo‘lda shakllantirilgan lingvistik bazadan ko‘ra kattaroq natija beradi. Taklif etilayotgan modelimizda (Word2Vec + PMI) 10 000 ta gapdan iborat korpusga ishlov berish va 5 000 ta so‘zlik lug‘atni generatsiya qilish jarayoni (dastlabki preprocessing va AI-labeling bosqichlarini hisobga olgan holda) zamonaviy hisoblash quvvatlarida (masalan, GPU bazasidagi tizimlar) bir necha daqiqadan oshmaydi $T_{auto} < 5\ min$.

Vaqt unumdorligi koeffitsiyentiniquyidagi formula orqali ifodalash mumkin:

$$K = \frac{T_{manual}}{T_{auto}} = \frac{5000}{5} = 1000$$

Bu shuni anglatadiki, taklif etilayotgan avtomatlashtirilgan tizim an’anaviy usuldan 1000 baravardan ortiqroq unumdorlikni ta’minlaydi.

O‘zbek tili uchun distributiv semantika va statistik vaznlashtirish algoritmlari asosida shakllantirilgan avtomatlashtirilgan lug‘at, inson omilini kamaytirgan holda katta hajmdagi ma’lumotlarni tahlil qilishda yuqori samaradorlikni ta’minlaydi. Mazkur tadqiqot natijalari kelgusida o‘zbek tilidagi ijtimoiy-siyosiy diskurslarni va foydalanuvchilar fikrini avtomatik monitoring qiluvchi intellektual tizimlarning asosi bo‘lib xizmat qiladi.

Foydalanilgan adabiyotlar ro‘yxati

1. Abdurakhmonova N., Shirinova R., Sayfullayeva R., Mengliev D., Ibragimov B., Ernazarova M. An annotated morphological dataset for Uzbek word forms: Towards rule-based and machine learning approaches // Data in Brief. – 2025. – Vol. 61. – Art. no. 111702. – DOI: 10.1016/j.dib.2025.111702.

2. Liu B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. – Cambridge, UK: Cambridge Univ. Press, 2015. – 384 p.

3. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. – Cambridge, MA, USA: MIT Press, 1999. – 680 p.

4. Mengliev D., Abdurakhmonova N., Barakhnin V., Vasliddinova K., Rahimov H., Djalolova K. Enhancing Sentiment Analysis in Uzbek Language Texts through Weighted Lexical Features // 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM). – Altai, Russian Federation, 2024. – Pp. 2450-2453. – DOI: 10.1109/EDM61683.2024.10615124.

5. Mengliyev B. et al. The morphological analysis and synthesis of word forms in the linguistic analyzer // Journal of Language and Linguistic Studies. – 2021. – Pp. 558-564.

6. Hamroyeva Sh. Morfologik analizatorni yaratish usullari // Uzbekistan language and culture. – 2022. – T. 5. – №. 1. – C. 88-108.