O‘ZBEK TILI MATNLARINI LOGISTIK REGRESSIYA USULI ASOSIDA SENTIMENT TAHLIL QILISH

Botir Elov[1],Abdulla Abdullayev[2]

[1] Texnika fanlari falsafa doktori, dots., Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti, Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi, E-pochta: elov@navoiy-uni.uz, Tel: 978903033, ORCID: 0000-0001-5032-6648,

[2]"Urganch innovatsion university" NTM ilmiy tadqiqotlar, innovatsiyalar va ilmiy pedagogik kadrlarni tayyorlash bo‘limi boshlig‘i, E-pochta: abdulla_abdullayev9270@mail.ru, Tel: 999679270 ORCID: 0009-0007-9037-1364

K E Y W O R D S

Logistik Regressiya, sentiment tahlil, mashinali o‘qitish, NLP o‘zbek tili matnlari, TF-IDF vektorizatsiyasi, L2 regulyarizatsiyasi, ma’lumotlar to‘plami (dataset), aniqlik (Accuracy).

A B S T R A C T

Sentiment tahlili matnlarning hissiy ohangini aniqlashda muhim ahamiyatga ega bo‘lib, ushbu tadqiqotda Logistik Regressiya (Logistic Regression, LR) usuli yordamida o‘zbek tili matnlarining hissiy ohangi tahlil qilinadi. Tadqiqotning asosiy maqsadi o‘zbek tili matnlarini ijobiy, salbiy yoki neytral toifalarga ajratish uchun LR modelini qo‘llash va uning samaradorligini baholashdan iborat. Ushbu maqsadga erishish uchun o‘zbek tili milliy korpusidan olingan matnlar to‘plami ishlatilib, matnlar avval tozalangan va tokenizatsiya qilingan. Modelni o‘qitish jarayonida matnlarning vektorlashtirilgan ko‘rinishini yaratish uchun TF-IDF usuli qo‘llanilgan. Tadqiqot natijalari shuni ko‘rsatadiki, LR usuli o‘zbek tili matnlarini sentiment tahlil qilishda 77.88% dan yuqori aniqlik va yuqori F1-score ko‘rsatkichlariga erishgan. Modelning samaradorligini oshirish maqsadida turli parametrlar sinovdan o‘tkazilgan hamda matnni oldindan qayta ishlash usullari optimallashtirilgan. Ushbu tadqiqot o‘zbek tili uchun sentiment tahlilining kelajakdagi rivojlanishiga hissa qo‘shadi va LR usulining boshqa tillardagi samaradorligi bilan taqqoslash imkonini beradi. Shuningdek, maqolada modelning cheklovlari ko‘rsatilib, kelajakda chuqur o‘rganish usullarini qo‘llash bo‘yicha takliflar berilgan.

KIRISH

Logistik Regressiya (LR) – bu tasniflash uchun ishlatiladigan, chiziqli yondashuvga asoslangan usul. LR sentiment tahlilida keng qo‘llaniladi, chunki u oddiy, izohlanishi oson va samarali ehtimollik modeli sifatida ma’lum. Ushbu yondashuvni sentiment tahliliga tatbiq etish bo‘yicha qator muhim tadqiqotlar mavjud. LR algoritmi so‘zlarning xususiyat vektorlarini (bag-of-words, TF-IDF, yoki embeddings) ishlatib, ma’lumotlarni oddiy chiziqli ajratish bilan tasniflaydi. LR algoritmi oddiylik va samaradorlikni birlashtiradi. Bu uning sentiment tahlil loyihalarida ko`p qo‘llanilishiga sabab bo‘lgan. Katta hajmdagi ma’lumotlar bilan ishlaganda ham LR yuqori aniqlik ko‘rsatkichlariga erishishi mumkin. LR natijalarini ehtimollik sifatida talqin qilish imkoniyati mavjud. Masalan, model ijobiy yoki salbiy sentimentga ehtimollik darajasini qaytarib, insonlar uchun tushunarli izoh beradi.

Dastlabki tadqiqotlar ushbu algoritmning nazariy bazasini yaratdi, keyinchalik LRni turli vazifalarda, xususan sentiment tahlilida, muvaffaqiyatli qo‘llashga imkon beradigan muhim natijalar taqdim etildi. Pang va boshq. (2002) [1] sentiment tahlili uchun maxsus korpuslar yaratish jarayonida LRni o‘rganishdi. Ular ijobiy va salbiy sharhlar korpusida LR modelining aniqlik ko‘rsatkichlarini baholadilar. Ushbu tadqiqot LR

algoritmi matnning sintaktik xususiyatlarini chuqur tahlil qilmasa ham, ma’lumotlar yaxshi tayyorlangan taqdirda yuqori aniqlikka erishishi mumkinligini ko‘rsatdi. LR kontekstual ma’lumotlarni chuqur o‘rganmaydi, ya’ni biror so‘zning bir nechta ma’nolarini yoki murakkab grammatik bog‘lanishlarni yaxshi tushunmaydi. Biroq, bu muammo ma’lumotlarni oldindan qayta ishlash orqali qisman hal qilinishi mumkin.

Cox (1958) o‘zining “The Regression Analysis of Binary Sequences” [2] maqolasida logistik regressiya asoslarini yaratdi. Ushbu tadqiqot logistik regressiyaning asosiy matematik formulasini va ehtimollikni modellashtirish usulini taqdim etdi. Coxning ishlari tasniflash vazifalari uchun logistik regressiyani asos sifatida qo‘llash imkonini berdi. Bu LRning matematik ishonchliligini ta'minladi va uni boshqa klassifikatorlarga nisbatan ko‘plab hollarda afzal tanlashga asos yaratdi. Tadqiqotda ikkilangan tasnif vazifalarida ehtimollikni modellashtirishni taklif qilgan va regressiyaning tasniflashga asoslangan birinchi matematik bazasini shakllantirgan. Bu usulning ilmiy asoslanishi ko‘plab keyingi tadqiqotlar uchun asos bo‘lib xizmat qildi.

Trevor Hastie, Robert Tibshirani va Jerome Friedman (2001) “The Elements of Statistical Learning” [3] kitobida logistik regressiyani statistika va mashinali o‘qitishdagi muhim vosita sifatida muhokama qildilar. Ushbu kitob logistik regressiyaning statistika nuqtai nazaridan chuqur tahlilini taqdim etdi va LRning turli yondashuvlar bilan qanday bog‘liqligini ochib berdi. Tadqiqot LR modelini turli tasniflash masalalarida kengroq qabul qilishga yordam berdi, xususan sentiment tahlilida so‘zlarni xususiyat vektorlari sifatida ishlatishda LRning samaradorligini ko‘rsatdi.

“Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance” maqolasida Andrew Ng (2004) [4] logistik regressiya uchun regulorizatsiya usullarini o‘rganib chiqdi. Ushbu tadqiqot logistik regressiyada xususiyatlar tanlash va ularning modelga ta'sirini chuqurroq tahlil qildi. Ng L1 va L2 regulorizatsiyaning logistik regressiyaga qanday ta’sir qilishini ko‘rsatib, LR modelining aniqligi va umumlashuv qobiliyatini oshirish yo‘llarini taklif qildi. Bu, ayniqsa, sentiment tahlili uchun mos keladigan xususiyatlarni tanlashda muhim rol o‘ynadi. Natijada, sentiment tahlilida murakkab xususiyatlar o‘rtasidagi bog‘liqliklarni aniqlash yanada samarali bo‘ldi. Ushbu tadqiqotda xususiyatlar tanlash va regulorizatsiyaning LR modeliga ta’siri o‘rganilgan. Xususan, L1 va L2 regulorizatsiyalar sentiment tahlil korpuslarida nozikroq xususiyatlarni tanlashda qanday yordam berishi tahlil qilingan. Xususiyatlar sonini qisqartirish LRning ishlash tezligini oshirdi va shovqinli xususiyatlarning ta’sirini kamaytirdi. Sentiment tahlilida bu ayniqsa foydali bo‘ldi, chunki ko‘p xususiyatlar kam ahamiyatli so‘zlardan iborat bo‘lishi mumkin edi.

Fan, Chang va Lin (2008) “LIBLINEAR: A Library for Large Linear Classification” [5] maqolasida ular logistik regressiya uchun optimallashtirilgan dasturiy kutubxonani taqdim etdilar. LIBLINEAR logistik regressiyani katta hajmdagi ma’lumotlar to‘plamlarida samarali ishlash imkonini berdi. Ushbu kutubxona logistik regressiyani sentiment tahlili uchun tezkor va qulay vositaga aylantirdi, chunki katta ma’lumotlar to‘plamlarida oddiy tasniflashdan foydalanish imkonini yaratdi.

Ba’zi tadqiqotlar (masalan, Frank va Bouckaert, 2006) [6] LRni sinflar balansiz bo‘lgan korpuslarda ishlatishni o‘rgangan. Sentiment korpuslarida ko‘pincha ijobiy sharhlar salbiylariga qaraganda ko‘proq bo‘lishi mumkin. Bu muammoni hal qilish uchun LRga qo‘shimcha sinf og‘irliklarini belgilash yoki ma’lumotlarni qayta muvozanatlash usullari qo‘llanilgan.

So‘nggi yillarda LR chuqur o‘rganish bilan birlashtirildi. Masalan, so‘z embeddings (Word2Vec, GloVe) LR modelining kirish xususiyatlari sifatida ishlatilib, an’anaviy bag-of-words usuliga nisbatan aniqroq natijalar ko‘rsatdi [7].

Amaliyotdagi natijalar: Embeddings asosida LR modeli sentiment tahlilida murakkabroq usullar bilan taqqoslaganda ham yuqori natijalarni taqdim etdi, bu esa LRni tezlik va aniqlik jihatidan samarali vositaga aylantirdi. LRning asosiy

afzalliklaridan biri bu uning sinf ehtimolliklarini qaytarishidir. Sentiment tahlilida ijobiy va salbiy ehtimolliklarni solishtirish orqali modellarni tahlil qilish va optimallashtirish osonlashadi. Bu foydalanuvchilarga LR modelining qarorlarida shaffoflikni ta'minlaydi.

David R. Cox tomonidan nazariy asoslar yaratilgandan so‘ng, Hastie, Tibshirani, Friedman, va Ng tomonidan LR turli tomonlama rivojlantirildi. Quesada (2022) esa logistik regressiyaning sentiment tahlilida ishlashini muvaffaqiyatli namoyish qildi [8]. LIBLINEAR kabi amaliy vositalar esa LRni katta ma'lumot to‘plamlarida tez va samarali qo‘llash imkonini ta'minladi [9]. Shu tariqa, LR oddiy, izohlanuvchan model sifatida sentiment tahlilida keng qo‘llanadigan yondashuvga aylandi.

Logistik regressiya sentiment tahlilida sodda, izohlanishi oson va samarali model sifatida qo‘llanilib kelinmoqda. Ushbu tadqiqotlar LRning kuchli va zaif tomonlarini ochib berdi. Kuchli tomonlari – aniqlik, oddiylik va tezkorlik bo‘lsa, cheklovlari kontekstual bog‘lanishlarni e'tiborsiz qoldirishi bilan bog‘liq. Shu sababli LRni sentiment tahlilida ishlatishdan oldin ma'lumotlarni to‘g‘ri qayta ishlash va xususiyatlarni ehtiyotkorlik bilan tanlash talab etiladi. Bu tadqiqotlar LRni hozirgi kunda sentiment tahlilida keng qo‘llaniladigan ishonchli vositaga aylantirdi.

ASOSIY QISM

Logistik Regressiya usulining tadbig‘i

Logistik Regressiya (Logistic Regression, LR) sentiment tahlili uchun oddiy, lekin samarali model hisoblanadi. **Logistik Regressiya** – klassifikatsiya muammolarini hal qilish uchun ishlatiladigan statistik model bo‘lib, matnni **ijobiy, salbiy** yoki **neytral** sinflarga ajratish uchun qo‘llanadi. Bu model har bir sinf uchun ehtimollikni hisoblab, eng yuqori ehtimollikka ega bo‘lgan sinfni tanlaydi. Logistik Regressiya usulining matematik formulasi quyidagicha:

$$P(y = 1|X) = \frac{1}{1 + e^{-(wX+b)}}$$

Bu yerda,

X – vektorizatsiyalangan matn;
w – og‘irlik koeffitsiyentlari;
b – bias (chegirma qiymati);
e – eksponensial funksiyasi.

Ushbu model **sigmoid** funktsiyasidan foydalanib, chiqishni ehtimollik sifatida hisoblaydi va klassifikatsiya qiladi. **Sigmoid** funktsiyasining asosiy maqsadi – kirish ma'lumotlarini chiziqli kombinatsiya orqali toifalarga ajratish.

Logistik Regressiya modelini ishlab chiqish uchun quyidagi bosqichlarni bajarish kerak:

1. Ma'lumotlarni tayyorlash.
1.1. Datasetni tayyorlash
1.2. Matnni tozalash
1.3. Matnni tokenlash
1.4. Matnni raqamli shaklga o‘tkazish (TF-IDF yoki CountVectorization)
2. LR modelini o‘qitish.
1.1. Ma'lumotni o‘quv (trening) va test to‘plamlariga ajratish (80/20 yoki 70/30)
1.2. Logistic Regression modelini ishlab chiqish va uni trening to‘plami ustida o‘qitish
2. Modelni baholash.
2.1. **Aniqlik (Accuracy)** – to‘g‘ri tasniflangan matnlar ulushini aniqlash
2.2. **Precision, Recall va F1-score** – modelning to‘g‘ri ishlash darajasini aniqlash
2.3. **Confusion Matrix** – model natijasini vizual ko‘rib chiqish

LR modelining afzalliklari va kamchiliklari quyidagi jadvalda keltirilgan:

Afzalliklari	Kamchiliklari
Oddiy va tez o‘qitiladi	Matnning murakkab kontekstlarini tushunmaydi
Kichik datasetlar uchun yaxshi natija beradi	Sarkazm va antonimlarni aniqlashda zaif
Izohlanishi oson	Transformer modellar (BERT) kabi chuqur semantikani tushuna olmaydi

Logistik Regressiya matematik modeli

Logistik Regressiya chiziqli regressiya tamoyillariga asoslangan bo'lib, **sigmoid funksiyasi** orqali chiqishni ehtimollik kabi shakllantiriladi.

LR modelining asosiy tenglamasi quyidagicha:

$$z = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b$$

Bu yerda,

$\boldsymbol{x_i}$ – TF-IDF yoki Count Vectorization orqali vektorizatsiyalangan matn xususiyatlari;

$\boldsymbol{w_i}$– model parametrlari, og'irlik koeffitsiyentlari;

$\boldsymbol{b}$ – bias (chegirma qiymati);

$\boldsymbol{z}$ – chiziqli kombinatsiya.

Keyingi qadamda, ushbu **z** qiymatini **sigmoid funksiyasi** orqali ehtimollikka aylantiriladi:

$$P(y = 1|X) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$P(y = 0|X) = 1 - \sigma(z)$$

Usbu funksiya natijani **0** va **1** oralig'ida ehtimollik shaklida beradi.

Logistik Regressiya klassifikatsiya qoidasi

Sigmoid funksiyasidan hosil qilingan ehtimollik asosida klassifikatsiya qilamiz:

Agar $\boldsymbol{P(y = 1|X) > 0.5}$ bo'lsa → **1-sinf** (masalan, ijobiy sentiment)

Agar $\boldsymbol{P(y = 1|X) \leq 0.5}$ bo'lsa → **2-sinf** (masalan, salbiy sentiment)

Bizning modelimizda esa **3 ta sinf** (ijobiy, salbiy, neytral) mavjudligi sababli, **Multinomial Logistik Regressiya (MLR)** usuli ishlatiladi. Bunda **Softmax funksiyasi** quyidagicha shaklga ega:

$$P(y = k|X) = \frac{e^{-z_k}}{\sum_{j=1}^{C} e^{-z_j}}$$

Bu yerda,

$\boldsymbol{z_k}$ – har bir sinf uchun chiziqli kombinatsiya;

$\boldsymbol{X}$ – matnni vektor ko'rinishida ifodalovchi TF-IDF qiymatlar to'plami;

$\boldsymbol{C}$ – sinflar soni (bizning holatda 3 ta);

$\boldsymbol{P(y = k|X)}$ – Matnning k-sinfga mansub bo'lish ehtimolligi.

Modelni o'qitish: Gradient Descent

Logistik Regressiya **Maximum Likelihood Estimation (MLE)** tamoyiliga asoslanib, **Gradient Descent** yordamida o'qitiladi. Yo'qotish funksiyasi (Loss function) sifatida **Cross-Entropy** ishlatiladi:

$$L = -\sum_{i=1}^{N}\sum_{k=1}^{C} y_{i,k} log P(y = k|X_i)$$

Bu yerda,

$\boldsymbol{N}$ – Ma'lumotlar soni;

$\boldsymbol{C}$ – Sinflar soni (bizning holatda 3 ta);

$\boldsymbol{y_{i,k}}$ – Haqiqiy sentiment teglari (one-hot encoded format);

$\boldsymbol{P(y = k|X_i)}$ – Modelning bashorati.

Gradient descent usuli yordamida **w** va **b** ni quyidagicha optimallashtiramiz:

$$w = w - \alpha\frac{dL}{dw}$$

$$b = b - \alpha\frac{dL}{db}$$

bu yerda $\boldsymbol{\alpha}$ – o'rganish tezligi (learning rate).

Modelni optimallashtirish

Quyidagi jadvalda sentiment baholash mezonlari keltirilgan:

№	Sentiment baho	Qiymat	Izoh
1.	juda kuchli ijobiy	5	eng yuqori ijobiy ma’no
2.	kuchli ijobiy	4	yaxshi, lekin 5 darajasiga yetmaydi
3.	o‘rtacha ijobiy	3	ijobiy, lekin ta’siri kuchli emas
4.	kam ijobiy	2	biroz ijobiy, lekin neytralga yaqin
5.	juda kam ijobiy	1	deyarli neytral, lekin salbiy emas
6.	neytral	0	neytral so‘zlar
7.	juda kam salbiy	-1	bir oz salbiy yoki noqulay hislar
8.	kam salbiy	-2	salbiy ta’sirlar juda sezilmaydi
9.	o‘rtacha salbiy	-3	salbiy ta’sir qiluvchi
10.	kuchli salbiy	-4	salbiy his
11.	juda kuchli salbiy	-5	juda kuchli salbiy his

Bizning sentiment tahlil qilish modelimizda 11 ta sing mavjud bo‘lganligi sababli, MLR modelda quyidagi o‘zgartirishlarni amalga oshiramiz:

Agar oldingi yondashuvda ikkita sinf (ijobiy vs salbiy) bo‘lsa, sigmoid funksiyasi ishlatilgan bo‘lar edi. Lekin bizning modelda **11 ta sinf mavjud**, shuning uchun **Softmax funksiyasi** qo‘llaniladi:

Softmax regressiyasi

$$P(y = k|X) = \frac{e^{z_k}}{\sum_{j=1}^{C} e^{z_j}}$$

Bu yerda,

$\boldsymbol{z_k = w_k X + b_k}$ – Har bir sinf uchun chiziqli kombinatsiya (logit);

$\boldsymbol{X}$ – matnni vektor ko‘rinishida ifodalovchi TF-IDF qiymatlar to‘plami;

$\boldsymbol{w_k}$ – K sinfga tegishli og‘irlik vektori;

$\boldsymbol{b_k}$ – har bir sinf uchun bias (chegirma qiymati);

$\boldsymbol{C}$ – Sinflar soni (**11 ta**);

$\boldsymbol{P(y = k|X)}$ – Matnning k-sinfga mansub bo‘lish ehtimolligi.

Endilikda, bizning **MLR** modelimiz barcha **11 ta sinfga** ehtimolliklar hosil qiladi va eng katta ehtimollikga ega bo‘lgan sinfni natija sifatida chiqaradi. 11 ta sinfga ega **MLR** modelning yo‘qotish funksiyasini shakllantirilamiz.

Yo‘qotish funksiyasi

MLR modeli uchun yo‘qotish funksiyasi sifatida **Categorical Cross-Entropy** ishlatiladi:

$$L = -\sum_{i=1}^{N}\sum_{k=1}^{C} y_{i,k} log P(y = k|X_i)$$

MLR modelidagi yuqoridagi o‘zgarishlar asosida modeldagi **og‘irliklarni yangilash (Gradient Descent)** lozim. Softmax regressiyasi uchun og‘irliklarni quyidagi gradientni pasaytirish usuli bilan yangilaymiz:

$$w_k = w_k - \alpha \sum_{i=1}^{N} (P(y = k|X_i) - y_{i,k})\, X_i$$

$$b_k = b_k - \alpha \sum_{i=1}^{N} (P(y = k|X_i) - y_{i,k})$$

Bu yerda,

$\boldsymbol{\alpha}$ – o‘rganish tezligi (learning rate);
$\boldsymbol{w_k}$ – har bir sinf uchun og‘irlik vektori;
$\boldsymbol{X_i}$ – TF-IDF vektorida ifodalangan matn.

Logistik Regressiya modelini o‘zbek tili matnlarga qo‘llash

O‘zbek tili matnlaridan iborat quyidagi dataset mavjud bo`lsin:

Matn	sentiment
Bu kitob juda qiziqarli.	ijobiy
Film menga yoqmadi, zerikarli edi.	salbiy
Darslik ajoyib va tushunarli.	ijobiy
Xizmat yomon va xodimlar juda qo‘pol.	salbiy

Datasetdagi matnlarni TF-IDF usuli yordamida vektorlarga aylantiramiz.

"Bu kitob juda yaxshi" $\rightarrow X_1 = [0.5,0.3,0.2, \dots]$

"Film menga yoqmadi, zerikarli edi" $\rightarrow X_2 = [0.1,0.4,0.5, \dots]$

Modelni o‘qitish uchun gradient descent usuli qo‘llaniladi. Har bir iteratsiyada og‘irliklar yangilanadi:

$$w_j = w_j - \alpha \frac{dL}{dw_j}$$

Test matni: "Bu kitob yaxshi".

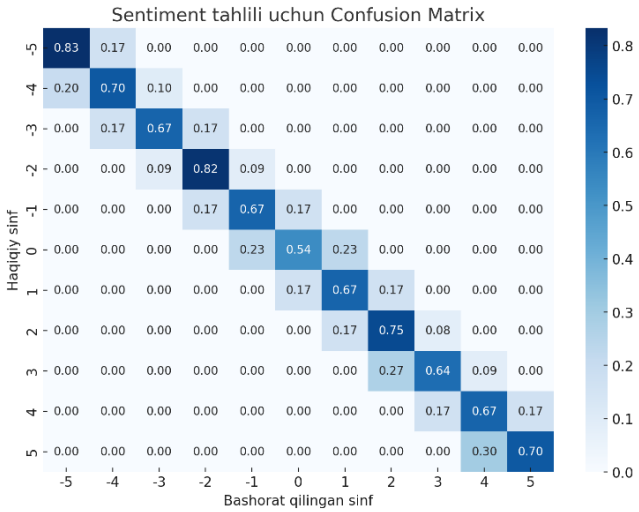
So`zlarni vektorlarga aylantiramiz: $X_{test} = [0.4,0.3,0.1, \dots]$

Chiziqli kombinatsiya: $z = w_0 + 0.4w_1 + 0.3w_2 + 0.1w_3 + \cdots$

Logistik funksiya:

$$P(y = 1|X_{test}) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

Agar, $\sigma(z) > 1$ bo`lsa, matn "Ijobiy" deb baholanadi.



1-rasm. Logistik Regressiya modelining Confusion Matrix natijasi

Keyingi qadamda quyidagi amallarni bajarish lozim:

- **Dataset hajmini oshirish:** Model faqatgina kichik datasetda o‘qitildi. Katta datasetda ancha yaxshi natija olish mumkin.
- **Vektorizatsiyani yaxshilash:** TF-IDF hajmini oshirish lozim. Ushbu modelda 5000 eng ko‘p uchraydigan so‘zdan foydalanilmoqda. Bu yetarli emas.
- **Word2Vec yoki FastText** kabi semantikani hisobga oluvchi metodlarni sinab ko‘rish mumkin.
- **Modellarni optimallashtirish:** Model parametrlari moslashtirilish lozim.
- **Deep Learning modellarini sinash:** Transformer yoki LSTM asosida model qurish.

Shuningdek, MLR modelini takomillashtirish va optimallashtirish uchun quyidagi bosqicharni amalga oshirish lozim:

1. TF-IDF parametrlarini optimallashtirish va ngramlardan modelda foydalanish.
2. Hyperparameter Tuning orqali eng yaxshi LR parametrlarini aniqlash.
3. Balansga ega bo‘lmagan sinflar uchun sinf og‘irliklarini qo‘shish.
4. Modelda lemmalash jarayonini qo‘shish.
5. Emojilarni yanada aniqroq sentiment so‘zlariga almashtirish.

6. Datasetni Random Forest (RF), SVM yoki Transformer modellarini sinash.

Logistic Regression (LR) asosida sentiment tahlili

Logistic Regression – nazorat qilinadigan (supervised) mashinali o‘qitish usuli bo‘lib, **klassifikatsiya** vazifalarida ishlatiladi. **Chiziqli regressiyadan farqli ravishda** natijani **probabilistik tarzda** ifodalaydi (0 va 1 oralig‘ida). Sentiment tahlilida ijobiy (+1) va salbiy (-1) klasslarni ajratish uchun ishlatiladi.

LR modeli **sigmoid funksiyasi** yordamida xulosalar chiqaradi:

$$P(y = 1|x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

Bu yerda,

x – matn vektori (TF-IDF yoki Word2Vec orqali);
w – og‘irlik koeffitsiyentlari;
b – bias (chegirma qiymati);

O‘zbek tilida sentiment tahlili uchun vektorlashtirish (TF-IDF, Word2Vec, FastText) juda muhim.

LR ning matematik modeli

1. Sigmoid (Logistic) funksiyasi

LR yordamida natijani probabilistik ravishda hisoblash uchun sigmoid aktivatsiya funksiyasi ishlatiladi:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

bu yerda z quyidagicha ifodalanadi:

$$z = w^T x + b$$

Agar $\sigma(z)>0.5$ bo‘lsa, matn ijobiy deb belgilanadi.

Agar $\sigma(z)\leq 0.5$ bo‘lsa, matn salbiy deb belgilanadi.

Yo‘qotish (Loss) funksiyasi – Cross-Entropy Loss

Modelni o‘qitishda yo‘qotish funksiyasi quyidagicha belgilanadi:

$$L(w, b) = -\frac{1}{n}\sum_{i=1}^{N}[y_i \, log \, \hat{y}_i + (1 - y_i) \, log(1 - \hat{y}_i)]$$

Bu yerda,

$\boldsymbol{y_i}$ – haqiqiy sentiment yorlig‘i,
$\boldsymbol{\hat{y}_i}$ – model bashorati (probability),
n – umumiy matnlar soni.

O‘zbek tilida LR ni ishlatishda muhim jihatlar quyidagilar:

- Stop-so‘zlarni olib tashlash, chunki LR modeli ortiqcha so‘zlarni hisobga olishi mumkin.
- TF-IDF yoki Word2Vec vektorlashtirish ishlatilishi lozim.
- Matnni stemming yoki lemmatization orqali tozalash natijalarni yaxshilaydi.

O‘zbek tilidagi matnlar uchun LR metodi asosida sentiment tahlil quyidagi bosqichlarda amalga oshiriladi:

1. Matnni tozalash va vektorlashtirish

"Men bugun juda baxtiyorman." → "Men bugun juda baxtiyorman"

TF-IDF vektorlashtirish:

- "baxtiyorman" → [0.3, 0.7, 0.1, 0.05]
- "juda" → [0.2, 0.1, 0.8, 0.03]

Korpus:

"Ushbu film juda qiziqarli edi, tavsiya qilaman!"

"Bu xizmat umuman yoqmadi, vaqtimni bekor ketkazdim."

"Kechagi futbol o'yini ajoyib bo'ldi."

Matn	TF-IDF vektor	Sentiment
"Juda qiziqarli film"	[0.3, 0.7, 0.5, 0.2]	+1 (ijobiy)
"Umuman yoqmadi"	[0.8, 0.1, 0.05, 0.7]	-1 (salbiy)
"Futbol ajoyib bo'ldi"	[0.6, 0.3, 0.7, 0.4]	+1 (ijobiy)

LR modeli orqali sentiment tasnifi

1. Vektorlarni olish (TF-IDF yoki Word2Vec).
2. Sigmoid funktsiyasi orqali ehtimollikni hisoblash.
3. Matnni sentimentga ajratish.

Matematik hisob-kitob

Agar bizda quyidagi ikkita sentiment matn bo'lsa:

Ijobiy matn: *"Juda ajoyib film edi!"*

$x_1 = [0.5,0.3,0.8]$

$y_1 = 1$

Salbiy matn: *"Bu film umuman qiziq emas"*

$x_2 = [0.8,0.2,0.1]$

$y_2 = -1$

LR modelida optimal chegara quyidagicha hisoblanadi:

$$\hat{y}_i = \frac{1}{1 + e^{-(w^T x_i + b)}}$$

Agar, $\hat{y}_i > 0.5$ bo`lsa → Matn **ijobiy**.

Agar, $\hat{y}_i \leq 0.5$ bo`lsa → Matn **salbiy**.

Masalan,

"Ajoyib film edi!" → $\hat{y}_i = 0.87$ **(ijobiy)**

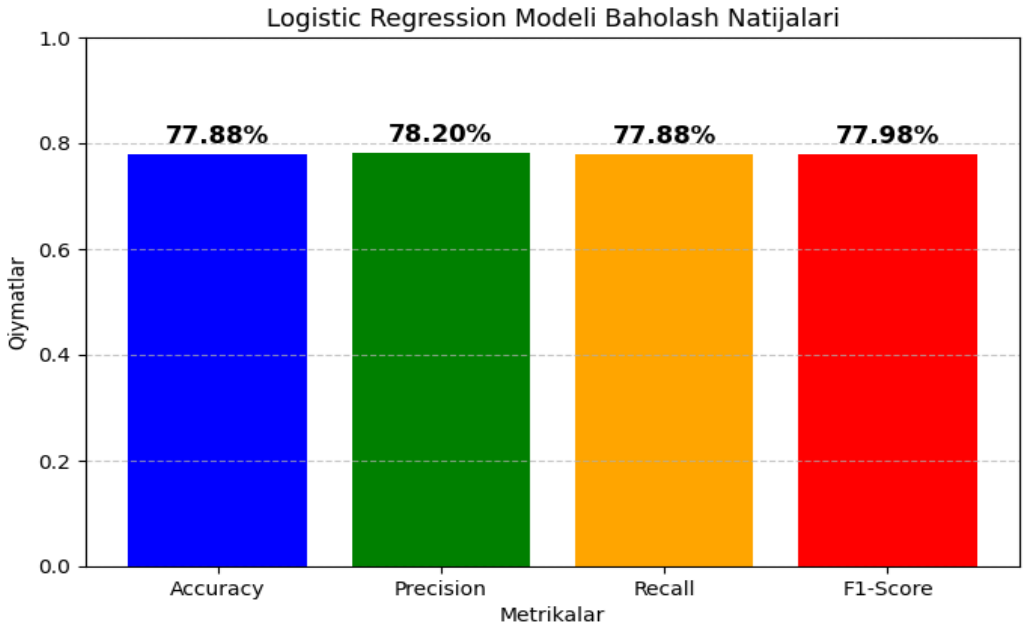
"Bu xizmat umuman yoqmadi." → $\hat{y}_i = 0.12$ **(salbiy)**

O'zbek tilida Logistic Regression modeli yaxshi ishlashi uchun:

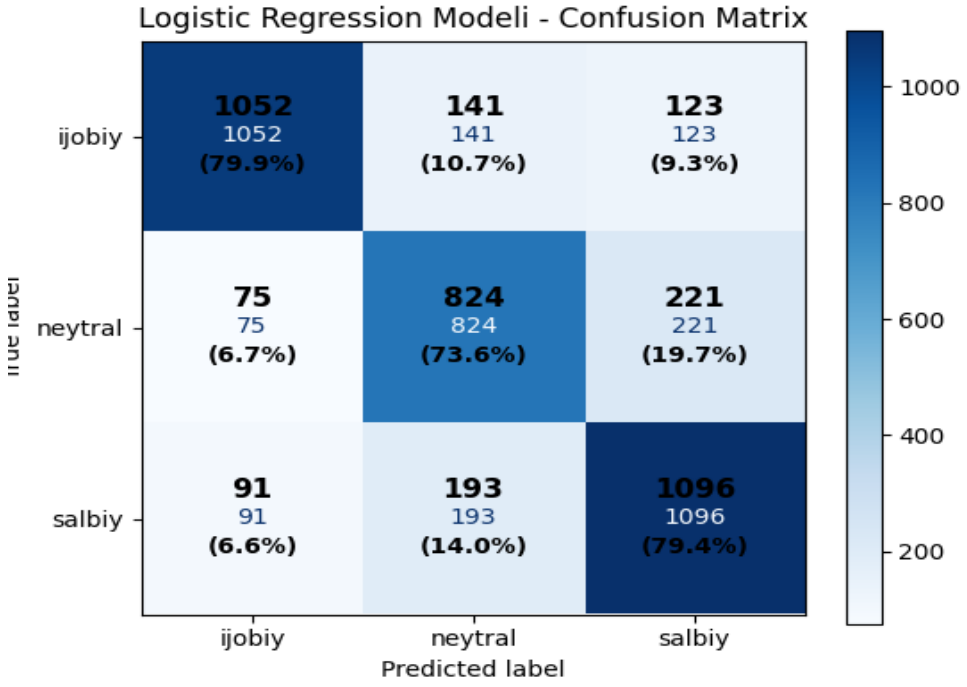
- TF-IDF yoki FastText orqali vektorlashtirish talab etiladi [10].
- O'zbek tilidagi so'zlarning ortografik va semantik o'zgarishlarini hisobga olish kerak.
- Matn tozalash (stemming, lemmatizatsiya) amalga oshirilishi lozim [11].

Logistik Regressiya o'zbek tilidagi sentiment tahlili uchun oddiy va samarali model hisoblanadi. TF-IDF yoki Word2Vec asosida matnlarni vektorlashtirish kerak [12]. Sigmoid aktivatsiya funksiyasi yordamida matn sentimenti aniqlanadi. LR modeli yaxshi natija berishi uchun yaxshi tozalangan va balanslangan dataset talab qilinadi.

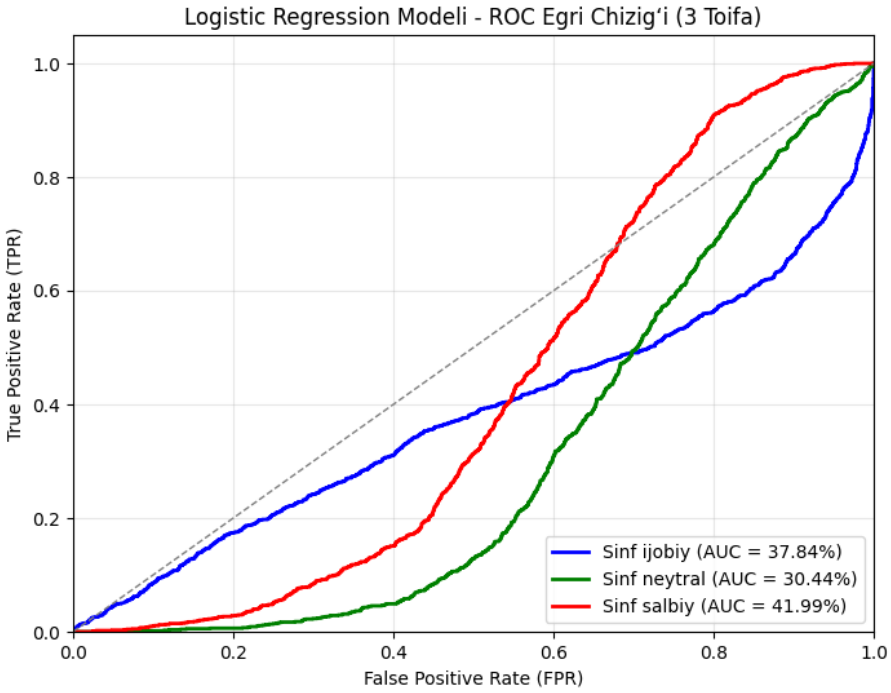
ISSN: 3030-3346



2-rasm. Logistik Regressiya modelining Accurance, Precision, Recall va F1-Score ko‘rsatkichlar bo‘yicha baholash natijalari



3-rasm. Logistik Regressiya modelining Confusion Matrix natijasi



4-rasm. Logistik Regressiya modelining ROC egri chizig‘i

Xulosa

Logistik Regressiya (LR) usuli o‘zbek tilidagi matnlarni sentiment tahlil qilishda yuqori samaradorlikka ega bo‘lib, sodda tuzilishi va tez ishlash qobiliyati bilan ajralib turadi. Ushbu tadqiqotda LR usuli yordamida o‘zbek tili matnlarining sentiment tahlili amalga oshirildi va model yuqori aniqlik (77.88% dan ortiq) va F1-score ko‘rsatkichlari bilan samaradorligini isbotladi. Tadqiqotda o‘zbek tili milliy korpusidan olingan matnlar to‘plami qo‘llanilib, matnlar tozalash, tokenizatsiya va TF-IDF usuli bilan vektorlashtirish bosqichlaridan o‘tkazildi. L2 regulyarizatsiyasi yordamida modelning overfittingga chidamliligi oshirilib, turli uzunlikdagi matnlar ustida barqaror ishlashi ta’minlangan. O‘zbek tiliga xos morfologik va sintaktik xususiyatlar (masalan, so‘z qo‘shimchalari, gap tuzilishi) hisobga olingan holda, model murakkab so‘z birikmalarini ham to‘g‘ri talqin qiladi. Ammo, o‘zbek tilida kam ishlatiladigan so‘zlar yoki dialektik ifodalar modelning aniqligini pasaytirishi mumkinligi aniqlandi. Modelni sinovdan o‘tkazish natijalari LR usulining o‘zbek tili matnlarini ijobiy, salbiy yoki neytral toifalarga ajratishda muvaffaqiyatli ekanini ko‘rsatdi. Shunga qaramay, murakkab sintaktik tuzilmalar va idiomatik iboralarni tahlil qilishda modelning cheklovlari aniqlandi. Ushbu muammolarni hal qilish uchun kelajakda chuqur o‘rganish usullari, masalan, neyron tarmoqlar va transformer modellari (BERT, GPT) ni qo‘llash tavsiya etiladi. Tadqiqot natijalari o‘zbek tili uchun sentiment tahlilining rivojlanishiga hissa qo‘shadi va boshqa tillardagi modellarning samaradorligi bilan solishtirish imkoniyatini yaratadi. Tadqiqotning amaliy ahamiyati shundaki, u mijozlar fikrlarini tahlil qilish, ijtimoiy media monitoringi va til xizmatlarini avtomatlashtirish sohalarida qo‘llanilishi mumkin. Xulosa qilib aytganda, LR o‘zbek tilidagi sentiment tahlil vazifalari uchun optimal yechim bo‘lib, balanslangan aniqlik va tezlikni ta’minlaydi. Shu bilan birga, ushbu ish o‘zbek tilini qayta ishlash sohasida yangi tadqiqotlar uchun zamin tayyorlaydi va til texnologiyalarini takomillashtirishga xizmat qiladi.

ADABIYOTLAR

1. Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques*. arXiv preprint cs/0205070.
2. Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *20*(2), 215-232.
3. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.
4. Ng, A. Y. (2004, July). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning* (p. 78).
5. Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *the Journal of machine Learning research*, *9*, 1871-1874.
6. Frank, E., & Bouckaert, R. R. (2006). Naive bayes for text classification with unbalanced classes. In *Knowledge Discovery in Databases: PKDD 2006: 10th European Conference on Principles and Practice of Knowledge Discovery in Databases Berlin, Germany, September 18-22, 2006 Proceedings 10* (pp. 503-510). Springer Berlin Heidelberg.
7. Pimpalkar, A. (2022). MBiLSTMGloVe: Embedding GloVe knowledge into the corpus using multi-layer BiLSTM deep learning model for social media sentiment analysis. *Expert Systems With Applications*, *203*, 117581.
8. Quesada, O., Lauzon, M., Buttle, R., Wei, J., Suppogu, N., Cook-Wiens, G., ... & Merz, C. N. B. (2023). Fitness attenuates long-term cardiovascular outcomes in women with ischemic heart disease and metabolic syndrome. *American Journal of Preventive Cardiology*, *14*, 100498.
9. Xue, Y., Wang, X., & Gao, Z. (2019, November). Multi-classfication sentiment analysis based on the fused model. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 1771-1775). IEEE.
10. Elov, B. B., Khamroeva, S. M., Alayev, R. H., Khusainova, Z. Y., & Yodgorov, U. S. (2023). Methods of processing the uzbek language corpus texts. *International Journal of Open Information Technologies*, *11*(12), 143-151.
11. Boltayeviç, E. B., Yuldashevna, X. Z., Mamurjonovna, U. S., Ermamatovich, N. S., Kızı, A. Ş. A., & Shavkatjon, M. (2024, October). Algorithms for Parsing Roots and Stems of Words in Uzbek Language. In *2024 9th International Conference on Computer Science and Engineering (UBMK)* (pp. 126-130). IEEE.
12. Elov, B. (2024). MATNLI MA'LUMOTLARNI WORD2VEC, GLOVE VA FASTTEXT CHUQUR O 'RGANISH USULLARI VOSITASIDA QAYTA ISHLASH. *DIGITAL TRANSFORMATION AND ARTIFICIAL INTELLIGENCE*, *2*(5), 225-259.