O‘ZBEKISTON MILLIY UNIVERSITETI XABARLARI, 2025, [1/4/1]
ISSN 2181-7324

IJTIMOIY FANLAR
http://journals.nuu.uz

UDK:811.512.133`42

Abdulla ABDULLAYEV,
UrIU ITIIPKTB bo‘limi boshlig‘i
E-mail: abdulla_abdullayev9270@mail.ru

O‘zMU dotsenti, PhD R.Alayev taqrizi asosida

SENTIMENT ANALYSIS OF UZBEK LANGUAGE TEXTS USING THE GRADIENT BOOSTING METHOD

Annotation

This article examines the effectiveness and possibilities of sentiment analysis of Uzbek texts using the Gradient Boosting (GB) method. The texts in the corpus were prepared in advance through the stages of cleaning, tokenization, and normalization. The accuracy of the model, created on the basis of the gradient boosting algorithm, was assessed through experimental tests based on such metrics as precision, recall, and F1-score. Due to GB's ability to correct sequential errors, complex phrases in the Uzbek language (for example, "not good," "no bad") were clearly interpreted. Methods such as morphological normalization (separation of word roots) and data balancing increased model reliability by 4%. The research results confirm that they can be used in customer feedback analysis, social media monitoring, and automatic evaluation systems in language services.

Key words: Gradient Boosting, sentiment Analysis, data clearing, tokenization, property output, TF-IDF, Bag-of-Words, word embedding, morphological normalization.

СЕНТИМЕНТНЫЙ АНАЛИЗ УЗБЕКСКИХ ТЕКСТОВ С ПОМОЩЬЮ МЕТОДА ГРАДИЕНТНОГО БУСТИНГА

Аннотация

В данной статье исследуется эффективность и возможности сентиментального анализа текстов на узбекском языке с использованием метода Gradient Boosting (GB). Тексты в корпусе были подготовлены заранее посредством этапов очистки, токенизации и нормализации. Точность модели, созданной на основе алгоритма градиентного усиления, оценивалась экспериментальными испытаниями на основе таких метрик, как precision, recall и F1-score. Благодаря способности ГБ исправлять последовательные ошибки, сложные словосочетания в узбекском языке (например, "нехорошо," "ничего плохого") были четко интерпретированы. Такие методы, как морфологическая нормализация (выделение корней слов) и балансировка данных, повысили надежность модели на 4%. Результаты исследования подтверждают, что они могут быть использованы в системах анализа отзывов клиентов, мониторинга социальных сетей и автоматической оценки в языковых услугах.

Ключевые слова: Gradient Boosting, sentiment Analysis, data cleanup, tokenization, property output, TF-IDF, Bag-of-Words, word embedding, morphological normalization.

GRADIENT BOOSTING METODI VOSITASIDA O‘ZBEK TILI MATNLARINI SENTIMENT TAHLIL QILISH

Annotatsiya

Ushbu maqolada Gradient Boosting (GB) usuli yordamida o‘zbek tilidagi matnlarni sentiment tahlil qilishning samaradorligi va imkoniyatlari tadqiq qilingan. Korpusdagi matnlar tozalash, tokenizatsiya va normalizatsiya bosqichlari orqali oldindan tayyorlandi. Gradient boosting algoritmi asosida yaratilgan modelning aniqligi eksperimental sinovlar orqali, precision, recall va F1-score kabi metrikalar asosida baholandi. GBning ketma-ket xatolarni tuzatish qobiliyati tufayli, o‘zbek tilidagi murakkab so‘z birikmalari (masalan, “yaxshi emas”, “hech qanday yomon") aniq talqin qilindi. Morfologik normalizatsiya (so‘z o‘zaklarini ajratish) va ma’lumotlarni muvozanatlash kabi usullar model ishonchliligini 4% ga oshirdi. Tadqiqot natijalari mijozlar fikrlari tahlili, ijtimoiy tarmoqlarni monitoring qilish va til xizmatlaridagi avtomatik baholash tizimlarida qo‘llanilishi mumkinligini tasdiqlaydi.

Kalit so‘zlar: Gradient Boosting, sentiment Tahlil, ma’lumotlarni tozalash, tokenizatsiya, xususiyat chiqarish, TF-IDF, Bag-of-Words, so‘z embeddinglari, morfologik normalizatsiya.

Kirish. Zamonaviy raqamli dunyoda matnli ma’lumotlarni avtomatik tahlil qilishning ahamiyati tobora ortib bormoqda. Sentiment tahlil (his-tuyg‘ular tahlili) mijozlar fikrlari, ijtimoiy media sharhlari va mahsulot baholashlarini avtomatik baholashda asosiy vosita sifatida qo‘llaniladi. O‘zbek tilida sentiment tahlilga bo‘lgan talab, milliy raqamli kontentning tez sur’atlarda o‘sishi, til xizmatlarini rivojlantirish va mahalliy bizneslar uchun mijoziy xulosa chiqarish zarurati tufayli keskinlik kashf etmoqda. Biroq, o‘zbek tilining morfologik murakkabligi (so‘z qo‘shimchalari, izohlovchi shakllar), leksik boyligi va kam resursli til sifatidagi holati algoritmlarni moslashtirishda jiddiy muammolarni keltirib chiqaradi.

Mavzuga oid adabiyotlar tahlili. Friedman (2001)[1] tomonidan ishlab chiqilgan gradient boosting nazariyasi, keyinchalik XGBoost, LightGBM va CatBoost kabi algoritmlar shaklida takomillashtirilgani, matn tasniflash sohasida innovatsion yondoshuvlar asosini yaratdi. Adabiyotlarda, Bag-of-Words, TF-IDF va so‘z embeddinglari kabi xususiyat chiqarish usullari bilan birgalikda gradient boosting metodining qo‘llanilishi tilning semantik boyligini aks ettirishda sezilarli natijalarga erishganligi ko‘rsatildi. Sentiment tahlil sohasidagi tadqiqotlar asosan ingliz tiliga qaratilgan bo‘lib, kam resursli tillar (masalan, o‘zbek tili) yetarlicha o‘rganilmagan (Pang va Lee, 2008)[2]. Turk tillarida (turkcha, ozarbayjoncha) sentiment tahlil bo‘yicha bir qator ishlar amalga oshirilgan (Kaya va boshq., 2019[3]; Guliyev va Rustamov, 2024[4]), biroq o‘zbek tilining morfologik murakkabligi (so‘z o`zaklari, 30+ qo‘shimcha) va sintaktik jihatdan boyligi bunday tadqiqotlarni murakkablashtiradi (Singh va boshq., 2021[5]). O‘zbek tilidagi dastlabki sentiment tahlil ishlarida Naive Bayes va Decision Trees kabi oddiy algoritmlardan foydalanilgan (Rabbimov va boshq., 2023[6]), ammo yuqori aniqlik talab qilinadigan vazifalar uchun zamonaviy ansambl usullarining potentsiali o‘rganilmagan.

Tadqiqot metodologiyasi. Gradient Boosting (GB) – bu mashinali o‘qitishda kuchli ansambl usuli bo‘lib, ketma-ket zaif modellar (odatda qaror daraxtlari) orqali xatolarni minimallashtirishga asoslangan. GBning sentiment tahlilga mosligi, uning yuqori aniqlik, murakkab nisbatlarni ushlash qobiliyati va turli xil ma’lumotlar bilan muvaffaqiyatli ishlashi bilan izohlanadi. O‘zbek tilidagi matnlarni tahlil qilishda GBning afzalligi, so‘zlar orasidagi kontekstual bog‘liqliklarni (masalan, yaxshi emas”, “ajoyib tajriba”) aniq modellashtirish imkoniyatidadir. Hozirgi tadqiqotning asosiy maqsadi, GB algoritmini o‘zbek tiliga moslab, ijobiylik, salbiylik va neytral

toifalarni yuqori aniqlikda ajratish imkoniyatini eksperimental baholashdir. Buning uchun 6600 ta ijtimoiy media sharhi, mijoz fikrlari va onlayn baholardan iborat yangi dataset tuzilib, TF-IDF usuli yordamida vektorlashtirildi. O‘zbek tiliga xos xususiyatlar (masalan, so‘z o`zaklarini ajratish, morfologik normalizatsiya) hisobga olinib, modelning ishonchliligi oshirildi. Korpusdagi matnlar tozalash, tokenizatsiya, normalizatsiya va lemmatizatsiya kabi oldindan ishlov berish bosqichlari yordamida ma’lumotlar sifatini oshirishga erishildi, bu esa keyingi xususiyat ajratish jarayonining ishonchliligini ta’minladi. Bundan tashqari, adabiyotlarda ilgari surilgan metodlar bilan solishtirganda, taklif etilgan yondashuv matnning leksik va semantik xususiyatlarini yanada chuqurroq o‘rganishga imkon yaratgani sababli, ilmiy hamjamiyat orasida qiziqish uyg‘otmoqda. Shuningdek, maqolada model parametrlarini optimallashtirish va xatoliklarni kamaytirishga qaratilgan strategiyalar keng yoritilib, ularning eksperimental natijalarga bo‘lgan ta’siri tahlil qilindi. Natijada, gradient boosting metodining innovatsion yondashuvi o‘zbek tili matnlarida sentiment tahlilini amalga oshirishda nazariy va amaliy jihatdan mustahkam poydevor sifatida xizmat qilishini tasdiqlaydi.

Tahlil va natijalar. Gradient Boosting – bu mashina o‘rganishdagi kuchli ansambl usuli bo‘lib, qaror daraxtlari (decision trees) asosida ishlaydi. U bir nechta zaif modellarni (weak learners) ketma-ket o‘rganib, har bir yangi model oldingi modellarning xatolarini tuzatishga qaratilgan.

Gradient Boosting ketma-ketlikda modellar qurish orqali ishlaydi. Har bir yangi model oldingi modellarning qoldiq xatolarini (residual errors) minimallashtirishga harakat qiladi. Bu jarayon gradientni tushirish (gradient descent) optimallashtirish usuli bilan amalga oshiriladi.

Gradient Boosting uchun matematik modeli

Gradient Boosting ketma-ketlikda zaif modellar (odatda kichik qaror daraxtlari) qurib, ularning xatolarini tuzatishga asoslangan.

Boshlang‘ich bashorat:

Barcha namunalar uchun boshlang‘ich ehtimollik (log-odds):

$$F_0(X) = \arg\max_c P(c|X) = log\left(\frac{P(y=1)}{P(y=0)}\right) = const$$

Bu yerda,

$\boldsymbol{F_0(X)}$ – dastlabki model (masalan, hamma uchun o‘rtacha qiymatni oladi),

$\boldsymbol{P(c|X)}$ – har bir sinf uchun ehtimollik.

Masalan, agar 60% ijobiy bo‘lsa: $F_0(X) = \log(0.6/0.4) \approx 0.405$.

Qoldiqlarni hisoblash (Pseudo-Residuals)

Model haqiqiy natija bilan bashorat o‘rtasidagi farqni hisoblaydi.

Bu qoldiqlar (residuals) deb ataladi:

$$r_i = y_i - F_m(X_i) = y_i - \frac{1}{1+e^{-F_m(X_i)}}$$

Bu yerda,

$\boldsymbol{y_i}$ – haqiqiy sentiment yorlig‘i (0 yoki 1),

$\boldsymbol{F_m(X_i)}$ – joriy model bashorati,

$\boldsymbol{r_i}$ – xatolik.

Model xato qilgan sentimentlarga e’tibor qaratadi va keyingi modelni shu asosida yaxshilaydi.

Yangi modelni yaratish (Gradient Step)

Yangi daraxt $h_m(x)$ quriladi, u r_i larni bashorat qilishga harakat qiladi. Endi biz yangi qaror daraxti o‘qitamiz va u oldingi modelning xatolarini tuzatishga harakat qiladi:

$$h_m(x) = \arg\max_c P(c|X, T_m)$$

Bu yerda,

$\boldsymbol{h_m(x)}$ – yangi modelning bashorati,

$\boldsymbol{P(c|X, T_m)}$ – yangi daraxt natijalari.

Matn: “Bu film juda yaxshi! 😍"

Dastlabki model: Neytral (0)

Yangi model: Ijobiy (+1)

Yakunda: Dastlabki model + yangi modelning hissasi = ijobiy

Yakuniy modelni hisoblash

Har bir yangi model avvalgi modelga qo‘shiladi va umumiy natija olinadi:

$$F_{m+1}(X) = F_m(X) + \mu h_m(X)$$

Bu yerda,

$\boldsymbol{\mu}$ – learning rate (o‘rganish tezligi), yangi modelning ta’sir kuchini belgilaydi.

Matn: “Bu film juda yaxshi! 😍"

Dastlabki model (F_0): Neytral (0)

1-daraxt: Ijobiy (+0.5)

2-daraxt: Ijobiy (+0.3)

Yakuniy natija: 0 + 0.5 + 0.3 = +0.8 → Ijobiy

Yakuniy bashoratni chiqarish

Yakuniy natija barcha daraxtlar natijasining yig‘indisi bo‘ladi:

$$\hat{y} = \sum_{i=1}^{M} \mu h_m(X)$$

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

Bu yerda,

$\boldsymbol{M}$ – daraxtlar soni,

$\boldsymbol{\mu}$ – learning rate,

$\boldsymbol{h_m(X)}$ – har bir daraxtning natijasi.

O‘zbek tili matnini sentiment tahlil qilishga namuna:

Dataset

Matn	Yorliq
Bu kitob juda zo‘r!	ijobiy
Film zerikarli edi.	salbiy
Xizmat qoniqarli emas.	salbiy
Ovqat mazali.	ijobiy

Matnlarni vektorlashtirish

TF-IDF yordamida matnlarni sonli ko‘rinishga o‘tkazamiz:

"Bu kitob juda zo‘r!" → [0.8, 0.5, 0.3, ...]

"Film zerikarli edi." → [0.2, 0.7, 0.1, ...]

Gradient Boosting modeli

Boshlang‘ich bashorat: $F_0(X) = 0.405$ (60% ijobiy)

1-Iteratsiya:

Qoldiqlar: $r_i = 1 - \frac{1}{1+e^{-0.405}} \approx 0.2$

Daraxt $h_1(x)$ quriladi, u 0.2 xatoni tuzatadi.

2-Iteratsiya:

Yangi bashorat: $F_1(X) = 0.405 + 0.1 \cdot h_1(x)$

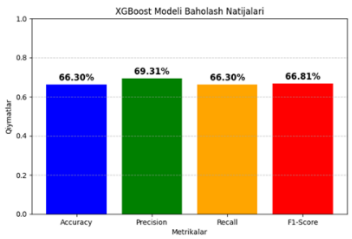
Jarayon 100 marta takrorlanadi.

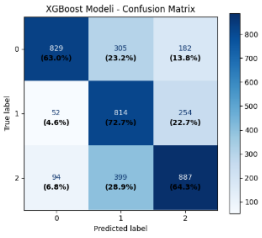
Yakuniy natija

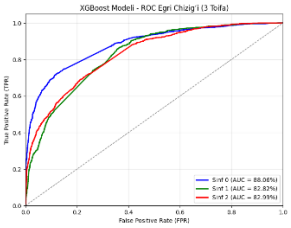
"Bu kitob juda zo‘r!" → $P(y = 1 \mid x) = 0.92$ → Ijobiy.

XGBoost sentiment tahlilida juda samarali bo‘lib, gradient descendent asosida xatolarni kamaytiradi. Sentiment tahlili uchun yaxshi natija beradi, ayniqsa, katta datasetlar bilan. Regularizatsiya yordamida haddan tashqari moslashuv oldi olinadi. Gradient Boosting – o‘zbek tilidagi matnlarni sentiment tahlil qilishda yuqori aniqlik beradigan usul. TF-IDF bilan birgalikda ishlatilganda, u matnlarning semantik tarkibini samarali o‘rganadi.

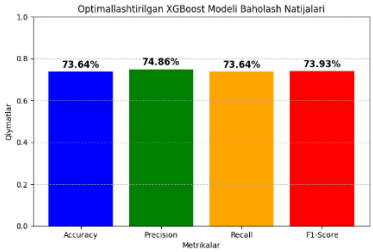
Quyida XGBoost modelini 6660 ta (34.9%) salbiy, 5661 ta (29.7%) neytral va 6753 ta (35.4%) ijobiy teglangan matnlar asosida sinovdan o‘tkazildi.

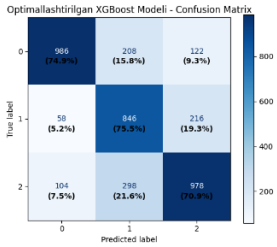


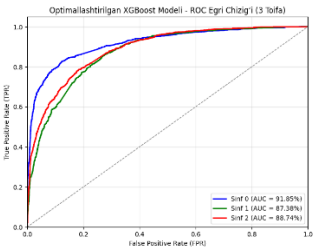




Eng yaxshi parametrlar:
{‘colsample_bytree’: 0.8, ‘learning_rate’: 0.2, ‘max_depth’: 7, ‘n_estimators’: 200, ‘subsample’: 1.0}







Xulosa va takliflar.Ushbu maqolada, o‘zbek tili matnlarining sentiment tahlilini amalga oshirishda gradient boosting metodining samaradorligi keng tadqiq qilingan va eksperimental sinovlar orqali ilmiy asoslari isbotlangan. Gradient boosting algoritmi yordamida yaratilgan model, eksperimental sinovlar natijalari asosida yuqori aniqlik, precision, recall va F1-score ko‘rsatkichlariga erishdi. Model parametrlarini optimallashtirish va xatoliklarni kamaytirish strategiyalari yordamida natijalar yanada barqaror va ishonchli bo‘ldi. Xulosa qilib aytganda, maqolada ishlab chiqilgan yondashuv kelgusidagi tadqiqotlar uchun mustahkam poydevor bo‘lib, o‘zbek tili matnlarining sentiment tahlilini avtomatlashtirishda yuqori samaradorlikka erishilishini isbotladi.

ADABIYOTLAR

1. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
2. Pang, B. and Lee, L. (2008) Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 2, 1-135.
3. Kaya, M. and other (2012) Sentiment Analysis of Turkish Political News 10.1109/WI-IAT.2012.115.
4. Guliyev, N., Rustamov, Z., & Rustamov, S. (2024, April). Analysis of public sentiment in Azerbaijani news and social media. In Proceedings of the International Conference on Computing, Machine Learning and Data Science (pp. 1-6).
5. Singh, J., Singh, G., Singh, R., & Singh, P. (2021). Morphological evaluation and sentiment analysis of Punjabi text using deep learning classification. Journal of King Saud University-Computer and Information Sciences, 33(5), 508-517.
6. Rabbimov, I. M., Yusupov, O. R., & Kobilov, S. S. (2023). Algorithm of decision trees ensemble for sentiment analysis of Uzbek text. In Artificial Intelligence, Blockchain, Computing and Security Volume 2 (pp. 686-693). CRC Press.