SENTIMENT TAHLILI UCHUN O‘RGATILGAN RESURLARI KAM TILLARDA MA’LUMOTLAR OMBORINI TUZISH

Allanazarova Saboxat Yusupboyevna,
Tayanch doktorant
allanazarova.sabosh@gmail.com
ToshDO‘TAU

Annotatsiya. Kam resursli tillar uchun sentiment tahlili tizimlarini rivojlantirish tabiiy tilni qayta ishlash (NLP) sohasidagi dolzarb muammolardan biridir. Ushbu maqolada kam resursli tillarda, xususan, o‘zbek tilida sentiment tahlili uchun lingvistik ta’minot yaratish jarayoni tahlil qilinadi. Leksik-semantik lug‘atlar, belgilangan korpuslar va annotatsiya usullarini ishlab chiqish, so‘z va iboralarning polaritesini aniqlash hamda ularni avtomatlashtirilgan tahlil tizimlariga integratsiya qilish usullari ko‘rib chiqiladi. Tadqiqotda o‘zbek tilining morfologik va sintaktik o‘ziga xosliklari hisobga olinib, mavjud xalqaro sentiment resurslari bilan taqqoslash amalga oshiriladi. Shuningdek, kam resursli tillar uchun lingvistik ta’minotni yaratishning asosiy muammolari va ularni bartaraf etish strategiyalari muhokama qilinadi. Ushbu ish natijalari o‘zbek tili uchun sifatli sentiment tahlili tizimlarini yaratishga hissa qo‘shishi bilan birga, kam resursli tillar uchun NLP infratuzilmasini rivojlantirishga ham xizmat qiladi.

Annotation. The development of sentiment analysis systems for low-resource languages is one of the critical challenges in the field of Natural Language Processing (NLP). This paper explores the process of creating linguistic resources for sentiment analysis in low-resource languages, specifically in Uzbek. It examines the development of lexicosemantic dictionaries, annotated corpora, and annotation methods, as well as automated techniques for determining the polarity of words and phrases, followed by their integration into sentiment analysis systems. The research takes into account the morphological and syntactic characteristics of the Uzbek language and compares it with existing international sentiment resources. Furthermore, the main challenges in creating linguistic resources for low-resource languages are discussed, along with strategies to overcome them. The findings of this study contribute to the development of high-quality sentiment analysis systems for the Uzbek language and support the advancement of NLP infrastructure for low-resource languages.

Аннотация. Развитие систем анализа сентимента для языков с ограниченными ресурсами является одной из актуальных проблем в области обработки естественного языка (NLP). В данной статье рассматривается процесс создания лингвистического обеспечения для анализа сентимента в языках с ограниченными ресурсами, в частности в узбекском языке. Исследуются методы разработки лексико-семантических словарей,

размеченных корпусов и аннотационных подходов, а также автоматизированные методы определения полярности слов и фраз с их последующей интеграцией в системы анализа сентимента. В работе учитываются морфологические и синтаксические особенности узбекского языка, проводится сравнительный анализ с международными сентиментными ресурсами. Кроме того, обсуждаются основные проблемы, возникающие при создании лингвистического обеспечения для языков с ограниченными ресурсами, а также предлагаются стратегии их преодоления. Полученные результаты способствуют созданию качественных систем анализа сентимента для узбекского языка, а также развитию NLP-инфраструктуры для языков с ограниченными ресурсами.

Kalit so‘zlar: *sentiment tahlili, kam resursli tillar, tabiiy tilni qayta ishlash, lingvistik resurslar, belgilangan korpus, leksik-semantik lug‘at, hissiy toifa, annotatsiya, NLP infratuzilmasi, avtomatlashtirilgan tahlil.*

1. KIRISH

So‘nggi yillarda tabiiy tilni qayta ishlash (NLP) va sun’iy intellekt texnologiyalarining rivojlanishi matnlarning avtomatlashtirilgan tahlilini keng qo‘llash imkonini bermoqda. Ushbu yo‘nalishlardan biri bo‘lgan sentiment tahlili matnlardagi ijobiy, salbiy yoki neytral kayfiyatni aniqlashga qaratilgan bo‘lib, ijtimoiy tarmoqlar, yangiliklar, sharhlar va boshqa matn manbalarini tahlil qilishda muhim ahamiyat kasb etadi. Biroq, sentiment tahlili tizimlarini yaratishda lingvistik resurslarning mavjudligi hal qiluvchi omillardan biri hisoblanadi. Ingliz, xitoy va boshqa ko‘p resursli tillar uchun keng ko‘lamli leksik-semantik lug‘atlar, belgilangan korpuslar va annotatsiya qilingan ma’lumotlar bazasi mavjud bo‘lsa-da, kam resursli tillarda, jumladan, o‘zbek tilida bunday infratuzilma yetarlicha rivojlanmagan.

Kam resursli tillar – tabiiy tilni qayta ishlash (NLP) va sun’iy intellekt tizimlari uchun yetarli darajada lingvistik resurslarga ega bo‘lmagan tillardir [1]. Ushbu tillarda belgilanadigan korpuslar, leksik-semantik lug‘atlar, annotatsiyalangan ma’lumotlar bazasi va yetarli miqdordagi o‘quv materiallari mavjud emasligi sababli, mashinani o‘rganish modellari samarali ishlashi qiyinlashadi. Kam resursli tillarga odatda rasmiy tillar sifatida kamroq qo‘llaniladigan yoki iqtisodiy va ilmiy sohalarda yetarli darajada raqamlashtirilmagan tillar kiradi.

Kam resursli tillarning xususiyatlari quyidagilarga qarab aniqlanadi:

Matn resurslarining cheklanganligi – onlayn va bosma manbalarda bunday tillarda ma’lumotlarning yetishmovchiligi.

Belgilanmagan korpuslar muammosi – annotatsiyalangan va belgilangan NLP ma’lumotlar bazasining yo‘qligi yoki juda kichik hajmda bo‘lishi.

Leksik va morfologik murakkablik – ayrim kam resursli tillar murakkab morfologiyaga ega bo‘lib, NLP modellarining samarali ishlashiga ta’sir ko‘rsatadi.

Ko‘p tilli NLP modellariga integratsiya qiyinligi – ko‘p hollarda bu tillar uchun maxsus NLP modellarining yo‘qligi sababli boshqa tillardan transfer o‘rganish usullari talab etiladi.

Masalan, o‘zbek tili NLP jamoatchiligi tomonidan kam resursli til sifatida baholanadi, chunki sentiment tahlili, mashinada tarjima va nutqni avtomatik tanib olish uchun katta hajmli va sifatli korpuslar yetarli emas. Ingliz, xitoy yoki nemis tillari bilan solishtirganda, o‘zbek tilida NLP modellarini o‘rgatish va qo‘llashda sezilarli qiyinchiliklar kuzatiladi. Shu sababli, kam resursli tillarda sentiment tahlili uchun lingvistik ta’minot yaratish dolzarb ilmiy muammolardan biri hisoblanadi.

Kam resursli tillarda sentiment tahlilining samaradorligini oshirish uchun maxsus lingvistik resurslarni ishlab chiqish zarur. Bu jarayon o‘z ichiga leksik-semantik lug‘atlar yaratish, belgilangan korpuslar tayyorlash, polarite va hissiyot darajalarini aniqlash kabi muhim bosqichlarni oladi. O‘zbek tilining morfologik va sintaktik jihatdan boy va murakkab tuzilishi sentiment tahlilini yanada qiyinlashtiradi. Shunday ekan, ushbu tadqiqotda o‘zbek tili misolida kam resursli tillar uchun sentiment tahlili tizimlarini rivojlantirishga qaratilgan lingvistik ta’minotni yaratish masalasi yoritiladi.

2. Adabiyotlar tahlili

So‘nggi yigirma yilda sentiment tahlili tabiiy tilni qayta ishlash (NLP) sohasining dolzarb yo‘nalishlaridan biriga aylandi va ushbu sohada turli nazariy hamda amaliy tadqiqotlar olib borildi. Xususan, sentiment tahlili uchun maxsus leksik resurslar va tasniflovchi modellar ko‘plab yirik tillar, jumladan, ingliz, rus va turk tillari uchun ishlab chiqilgan. Biroq, kam resursli tillar, ayniqsa, agglutinativ tuzilishga ega bo‘lgan o‘zbek tili uchun bunday imkoniyatlar cheklangan bo‘lib, mavjud metod va resurslarning bevosita qo‘llanishi samarasiz bo‘lishi mumkin. Ayrim yondashuvlar sentiment lug‘atlarini tarjima qilish orqali foydalanishni taklif qiladi, ammo tillar orasidagi semantik, sintaktik va pragmatik tafovutlar natijasida bunday yondashuv yetarlicha aniqlikni ta’minlay olmaydi. Shu bois, har bir tilning morfologik va sintaktik o‘ziga xosliklarini inobatga olgan holda, maxsus sentiment lug‘atlari hamda tasniflash modellarini ishlab chiqish zarurati ortib bormoqda.

Rus tilida sentiment tahliliga oid tadqiqotlar lingvistik yondashuvlardan tortib, mashinani o‘rganish va chuqur neyron tarmoqlar asosidagi modellargacha bo‘lgan keng doirani qamrab oladi. Ayniqsa, oxirgi yillarda transformer modellar va kontekstga bog‘liq tahlil usullarining rivojlanishi sentimentni yanada aniqroq

aniqlashga imkoniyat yaratmoqda. O‘zbek tili uchun sentiment tahlili borasida esa hali ham yetarli tadqiqotlar mavjud emas, ammo so‘nggi yillarda ayrim ilmiy izlanishlar boshlangan.

D.B. Mengliyev boshchiligidagi tadqiqotchilar guruhi o‘zbek tilidagi sentiment tahlilini amalga oshirish uchun maxsus algoritm ishlab chiqib, har bir so‘z, ibora va frazeologizmlarga -3 dan +3 gacha bo‘lgan sentiment ballarini belgilash mexanizmini taklif etishadi[2]. Sinov natijalari ushbu algoritmning aniqligi 85–95% oralig‘ida ekanligini ko‘rsatgan, biroq tadqiqot davomida so‘z shakllarini yanada mukammal qayta ishlash va lug‘at bazasini kengaytirish zarurati aniqlangan.

E.Y. Akhmedov o‘z tadqiqotida sentiment tahlili jarayonida o‘zbek tilining leksik, grammatik va kontekstual xususiyatlarini hisobga olish lozimligini ta’kidlaydi. Xususan, chuqur o‘rganish modellaridan—LSTM, CNN va mBERT kabi ilg‘or yondashuvlardan foydalanish o‘zbek tilidagi sentiment tahlilining samaradorligini oshirish imkonini berishi ko‘rsatilgan[3].

K. Niyozmetova tomonidan olib borilgan tadqiqotda Toshkent shahridagi restoranlarga oid Google Maps’dagi foydalanuvchilar fikrlari tahlil qilingan. Tadqiqotchilar sentiment tasnifi uchun logistic regression modelidan foydalangan bo‘lib, oldindan qayta ishlash bosqichida stemming kabi agglutinativ tillar uchun muhim bo‘lgan usullarning samaradorligi kuzatilgan. Tadqiqot natijalari sentiment tahlili xizmat sifatini baholashda va iste’molchi sadoqatini oshirishda samarali vosita ekanligini tasdiqlaydi[4].

S. Matlatipov tadqiqotida o‘zbek tilidagi restoran sharhlaridan iborat sentiment datasetini yaratish va tahlil qilish jarayoni yoritilgan. O‘zbek tilining kam resursga ega tillardan biri ekani inobatga olinib, logistic regression, support vector machines (SVM) hamda chuqur o‘rganish modellari (RNN, CNN) sinovdan o‘tkazilgan. Ma’lumotlarni yig‘ish, oldindan qayta ishlash va baholash bosqichlari batafsil tahlil qilingan bo‘lib, stemming kabi metodlarning qo‘llanilishi natijalarning yaxshilanishiga xizmat qilgan. Eng samarali model 91% aniqlikka erishgan[5].

E. Kuriyozov tadqiqotida o‘zbek tilida ko‘p-toifali yangiliklar tasnifi uchun dataset yaratish va modellarni baholash jarayoni o‘rganilgan. Shu maqsadda 10 ta yangilik saytidan ma’lumotlar yig‘ilib, 15 toifani qamrab oluvchi dataset shakllantirilgan. Sinov natijalariga ko‘ra, RNN va CNN qoidaga asoslangan usullarga nisbatan samaraliroq ekanligi kuzatilgan, eng yuqori natijalar esa BERTbek modelida qayd etilgan. Ushbu natijalar o‘zbek tilidagi matn tasniflash bo‘yicha tadqiqotlar uchun muhim asos bo‘lib xizmat qilishi mumkin[6].

I. Rabbimovning tadqiqotida o‘zbek tilida ko‘p-toifali tasniflash amalga oshirilib, “Daryo” yangiliklar saytining 10 toifasiga tegishli dataset yaratilgan. Tadqiqotchilar SVM, DTC, RF, LR va MNB kabi mashinani o‘rganish modellaridan foydalanib, sentiment tasnifini amalga oshirishgan. Belgilar ajratishda TF-IDF va n-

gram metodlari qo‘llanilgan. 5-fold cross-validation orqali hiperparametrlar optimallashtirilgan va natijada 86.88% aniqlikka erishilgan. Ushbu tadqiqot natijalari o‘zbek tilidagi matnlarni avtomatik tasniflash uchun samarali usullarni ko‘rsatadi[7].

A. Yusufu va hamkor tadqiqotchilar guruhi kam resursli til sifatida o‘zbek tilida sentiment tahlilini o‘rganish maqsadida Google Play Store’dan 27,985 ta o‘zbekcha sharh yig‘ib, User Experience Honeycomb modelining olti asosiy komponenti bo‘yicha 43,712 jumlani annotatsiya qilishgan. Sentiment qutblanishini baholash uchun oldindan o‘qitilgan modellar hamda GCN asosida integratsiyalangan yondashuv ishlab chiqilgan. Multi-klassifikatsiya uchun F1 ko‘rsatkichida 0.30, binar tasniflash uchun esa 0.43 yaxshilanish kuzatilgan. Ushbu tadqiqot kam resursli tillarda foydalanuvchi tajribasini chuqur tahlil qilish uchun istiqbolli yondashuvlarni taqdim etadi[8].

Mazkur tadqiqotlar shuni ko‘rsatadiki, o‘zbek tilida sentiment tahlilini rivojlantirish uchun lingvistik resurslarni kengaytirish, maxsus sentiment lug‘atlarini ishlab chiqish hamda chuqur o‘rganish modellarini tatbiq etish muhim ahamiyatga ega. Ushbu yo‘nalishda tadqiqotlarni davom ettirish va NLP texnologiyalarini takomillashtirish kelajakda samarali natijalarga olib kelishi mumkin.

3. Metodologiya

Kam resursli tillar uchun sentiment tahlili tizimini yaratishda lingvistik ta’minotning yetishmovchiligi asosiy muammolardan biri hisoblanadi. Ushbu tadqiqotda o‘zbek tilida sentiment tahlili uchun zarur bo‘lgan lingvistik resurslarni yaratish metodologiyasi ishlab chiqildi. Metodologiya uch asosiy bosqichni o‘z ichiga oladi: (1) belgilangan korpus yaratish va annotatsiyalash, (2) leksik-semantik lug‘at ishlab chiqish, (3) sentiment tahlili modellarini sinovdan o‘tkazish va baholash. Baholash yarim avtomatik tarzda amalga oshirildi.

Belgilangan korpus yaratish va annotatsiyalash

Belgilangan korpus sentiment tahlili modelini o‘rgatish va baholash uchun asosiy manba hisoblanadi. O‘zbek tilida bunday korpuslarning mavjud emasligi sababli, qo‘lda annotatsiyalash va yarim-avtomatlashtirilgan yondashuvlardan foydalanish zarur. Annotatsiya jarayonida quyidagi xalqaro amaliy tajribalar tahlil qilindi:

- **SentiWordNet** (Esuli & Sebastiani, 2006)
- **SenticNet** (Cambria va boshqalar, 2012)

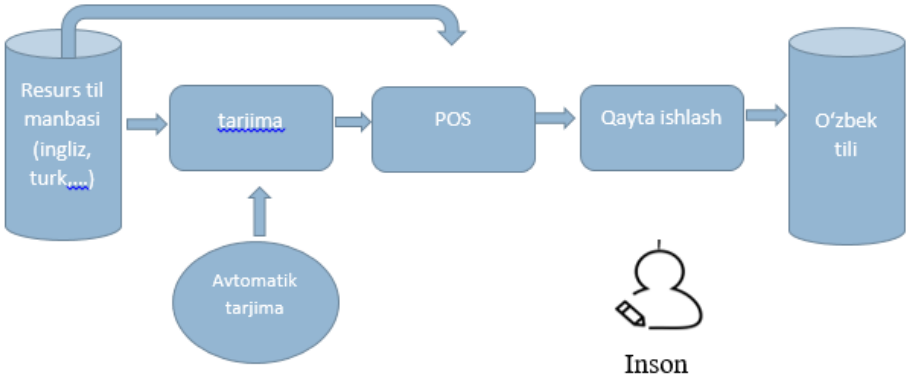
SentiWordNet, SenticNet score lari kiritildi. Bularni hammasidan foydalanib SentiUzNet lingvistik ta’minoti uchun pos, neg, obj score lari mashinaviy o‘qitish algoritmlari yordamida qo‘yildi.

O‘zbek tili uchun belgilangan korpus yaratishda ushbu ishlardan ilhomlanib, **ijtimoiy tarmoqlar (Twitter, Facebook), yangiliklar portallari (Kun.uz, Daryo.uz) va blog postlari** kabi turli manbalardan matnlar yig‘ildi. Tanlangan matnlar qo‘lda annotatsiyalanib, ijobiy, salbiy va betaraf toifalarga ajratildi. 5 nafar annotatorlar o‘rtasidagi kelishuv darajasini baholash uchun Cohen’s Kappa koeffitsienti (Cohen, 1960) ishlatildi. Belgilangan Cohen's kappa toifalar mosligi 0.79 ni ko‘rsatdi.

Leksik-semantik lug‘at ishlab chiqish

Leksik-semantik lug‘atlar sentiment tahlilining muhim komponenti hisoblanadi. Ushbu bosqichda o‘zbek tilida foydalanish mumkin bo‘lgan lug‘at yaratish uchun quyidagi usullardan foydalanildi:

Tarjima asosida lug‘at yaratish: SentiTurkNet lug‘atidagi fizika, kimyo, tibbiyot sohalariga oid 500 dan ortiq terminlar o‘zbek tiliga tarjima qilindi va ularning polaritesi lingvist mutaxassislar tomonidan tekshirildi.

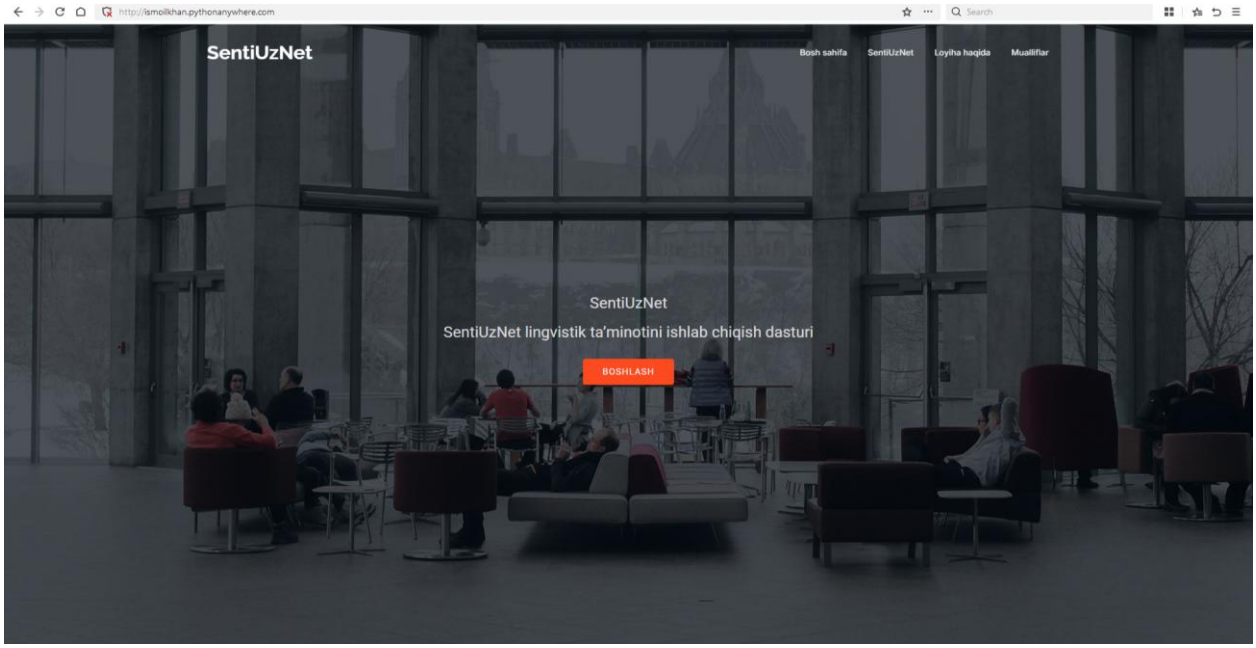


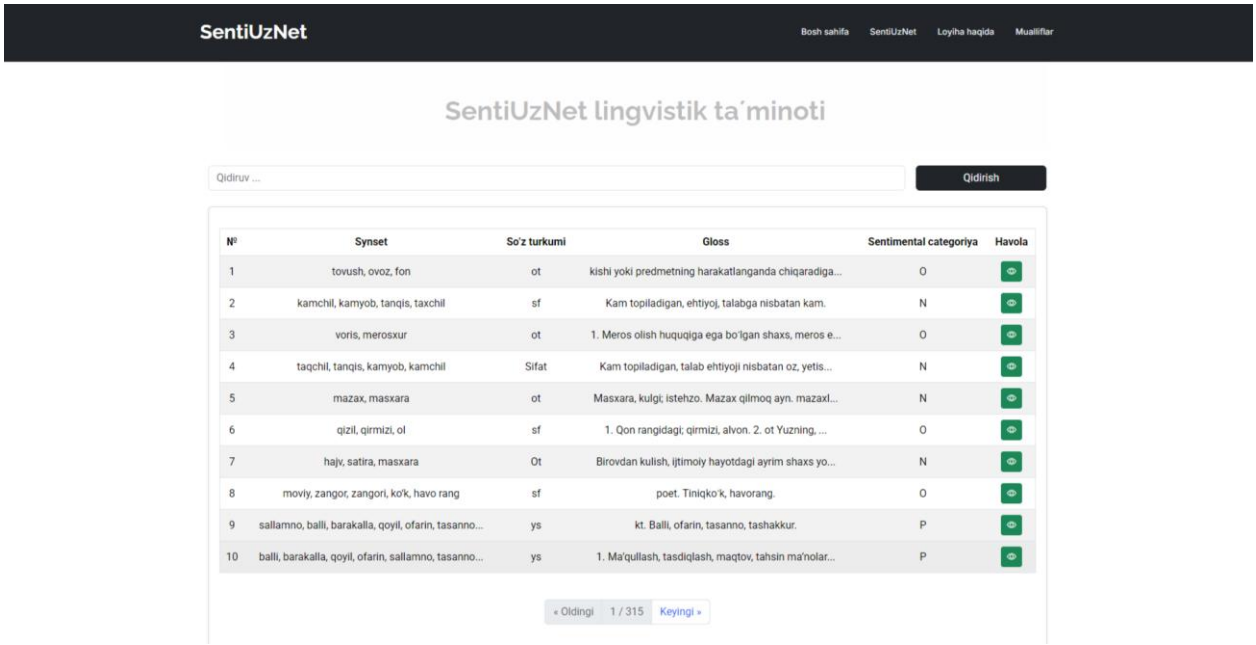
1- rasm. Tarjimaga asoslangan usulning bosqichlari.

O‘zbek tili uchun ishlab chiqilgan sentiment lug‘at ingliz (SentiWordNet), (SenticNet) har bir tillar orasidagi hissiy tasniflash farqlari statistik tahlil qilindi.

NATIJALAR

Har bir so‘z, sinset, gloss uchun PosPMI, NegPMI hisoblab chiqildi. Bunda o‘zbek tilidagi barcha sentimental tahlil datasetlaridan foydalanildi, jami 384 mingta sharh (review) yordamida hisoblandi.





O‘zbek tilida sentiment tahlilini baholash natijalari boshqa tillar uchun erishilgan natijalar bilan taqqoslandi. Masalan, O‘zbek tili uchun ishlab chiqilgan sentiment lug‘at ingliz tilidagi (SentiWordNet), (SenticNet) lug‘atlari orasidagi hissiy tasniflash farqlari statistik tahlil qilindi. Jami 11 ta ML algoritmi ishlatilgan bo’lib, 71-91% aniqlik ko‘rsatgan.

Model	Accuracy	Precision	Recall	F-Measure	Training Time (s)	Test Time (s)
Random Forest	0,9141	0,9137	0,9141	0,9139	1,5148	0,0090
XGBoost	0,9141	0,9137	0,9141	0,9138	0,9549	0,0060
LightGBM	0,9094	0,9089	0,9094	0,9091	0,8835	0,0040
Stacking Classifier	0,9094	0,9088	0,9094	0,9090	12,8707	0,1130
Logistic Regression	0,8951	0,8950	0,8951	0,8949	0,2822	0,0000
Linear Discriminant Analysis	0,8871	0,8864	0,8871	0,8865	0,0269	0,0010
Naive Bayes	0,8808	0,8853	0,8808	0,8811	0,0170	0,0010
Neural Network	0,8776	0,8787	0,8776	0,8779	3,2221	0,0010
SVM (SMO)	0,8760	0,8798	0,8760	0,8730	2,8455	0,0660
Decision Tree	0,8665	0,8659	0,8665	0,8650	0,0840	0,0010
K-Nearest Neighbors	0,7059	0,7042	0,7059	0,7022	0,1912	0,0191

Xulosa

Ushbu tadqiqot kam resursli tillar, xususan, o‘zbek tili uchun sentiment tahlili tizimlarini rivojlantirishning asosiy jihatlarini yoritdi. Tadqiqot natijalari shuni ko‘rsatadiki, sentiment tahlili uchun maxsus leksik-semantik lug‘atlar, belgilangan korpuslar va avtomatlashtirilgan annotatsiya tizimlari zarur. O‘zbek tilining agglutinativ tuzilishi va murakkab morfologiyasi mavjud modellarni to‘g‘ridan-to‘g‘ri qo‘llashda qiyinchiliklar tug‘dirishi aniqlangan. Ushbu tadqiqot natijalari nafaqat o‘zbek tilida, balki boshqa kam resursli tillar uchun ham NLP infratuzilmasini rivojlantirishga hissa qo‘shadi.

Foydalanilgan adabiyotlar:

1. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of ACL*, 93-103. https://doi.org/10.18653/v1/2020.acl-main.560
2. Mengliev, D.B., Abdurakhmonova, N.Z., Barakhnin, V.B., Vasliddinova, K., Rahimov, H., & Djalolova, K. (2024). 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), 2450-2453.
3. Akhmedov, E.Y., Palchunov, D.E., Khaitboeva, D.Z., Ibragimov, M.F., Sultanov, O.R., & Rakhimova, L.S. (2024). Sentiment Analysis in Uzbek Language Texts: a Study Using Neural Networks and Algorithms. 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), 2460-2464.
4. Kumushoy, N., Farogat, Y., & Ogabek, S. (2024). Sentiment Analysis In Uzbek Texts In The Restaurant Field. *2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE)*, 1470-1473.
5. Matlatipov, S., Rahimboeva, H., Rajabov, J., & Kuriyozov, E. (2022). Uzbek Sentiment Analysis Based on Local Restaurant Reviews. *ALTNLP*.
6. Kuriyozov, E., Salaev, U., Matlatipov, S., & Matlatipov, G. (2023). Text classification dataset and analysis for Uzbek language. *ArXiv, abs/2302.14494*.

7. Rabbimov, I., Kobilov, S.S., & Mporas, I. (2021). Opinion Classification via Word and Emoji Embedding Models with LSTM. *International Conference on Speech and Computer*.

8. **Yusufu, A., Ainiwaer, A., Li, B., Li, F., Yusufu, A., & Ji, H.** (2025). *Analysis of user experience in low-resource languages: A case study of the Uzbek language Google Play reviews*. **Information Processing and Management, 62**, 104015. https://doi.org/10.1016/j.ipm.2024.104015

9. **Cohen, J.** (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

10. Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. *Proceedings of LREC*, 417–422.

11. Cambria, E., Havasi, C., & Hussain, A. (2012). SenticNet 2. In *Proc. AAAI IFAI RSC-12*.