QORAQALPOQ TILIDAGI MATNLARNI MASHINALI O‘QITISH ALGORITMLARI YORDAMIDA TASNIFLASH

Jalelova Malika Moyatdin qizi
Berdaq nomidagi Qoraqalpoq davlat universiteti dotsent v.b.
Komekova Gulzira Sharbayevna
Berdaq nomidagi Qoraqalpoq davlat universiteti assistenti

Annotatsiya. Mazkur ishda qoraqalpoq tilidagi yangilik matnlarini avtomatik tasniflash masalasi ko‘rib chiqilgan. Tadqiqot doirasida qoraqalpoq tilida faoliyat yurituvchi internet manbalaridan yangilik maqolalari to‘plangan va maxsus korpus shakllantirilgan. Matnlarni tasniflash vazifasini yechish uchun Linear SVM, Logistik regressiya, Multinomial Naive Bayes, Random Forest va K-Nearest Neighbors algoritmlaridan foydalanilgan va ular orasida Random Forest modeli eng yuqori aniqlikni ko‘rsatgan. Olingan natijalar resurs-cheklangan tillar uchun klassik mashinali o‘qitish usullari samarali yechim bo‘la olishini tasdiqlagan.

Kalit so‘zlar: qoraqalpoq tili, matn klassifikatsiyasi, mashinali o‘qitish, TF-IDF, Random Forest.

Abstract. This paper addresses the problem of automatic classification of news texts in the Karakalpak language. Within the scope of the study, news articles were collected from online sources operating in the Karakalpak language, and a dedicated text corpus was constructed. To solve the text classification task, Linear SVM, Logistic Regression, Multinomial Naive Bayes, Random Forest, and K-Nearest Neighbors algorithms were applied, among which the Random Forest model demonstrated the highest accuracy. The obtained results confirmed that classical machine learning methods can serve as an effective solution for low-resource languages.

Keywords: Karakalpak language, text classification, machine learning, TF-IDF, Random Forest.

Аннотация. В данной работе рассмотрена задача автоматической классификации новостных текстов на каракалпакском языке. В рамках исследования были собраны новостные статьи из интернет-источников, функционирующих на каракалпакском языке, и сформирован специализированный корпус текстов. Для решения задачи классификации были использованы алгоритмы Linear SVM, логистической регрессии, Multinomial Naive Bayes, Random Forest и K-Nearest Neighbors, среди которых модель Random Forest показала наивысшую точность. Полученные результаты подтвердили, что классические методы машинного обучения могут выступать эффективным решением для языков с ограниченными ресурсами.

ISSN 3030-3907

Ключевые слова: каракалпакский язык, классификация текстов, машинное обучение, TF-IDF, Random Forest.

Kirish

Zamonaviy axborot asrida raqamli kontentning jadal o‘sishi butun dunyo tillarida matnli ma’lumotlarni avtomatik tahlil qilish zaruratini keltirib chiqarmoqda [1]. Biroq, bu imkoniyatlar asosan ingliz, xitoy, ispan kabi keng tarqalgan tillarda rivojlangan bo‘lib, kam resursli tillar uchun sun’iy intellekt texnologiyalarini qo‘llanilishi hali ham jiddiy ilmiy va amaliy muammo bo‘lib qolmoqda. Qoraqalpoq tili qoraqalpoq xalqining rasmiy tili bo‘lishiga qaramay, zamonaviy kompyuter lingvistikasi sohasida deyarli o‘rganilmagan tillardan biri hisoblanadi.

So‘nggi yillarda qoraqalpoq tilidagi raqamli kontentni ko‘payishi kuzatilmoqda. Karalpak.uz, qalpaq.info, erkin-qaraqalpaqstan.uz kabi yangilik portallari, ijtimoiy tarmoqlardagi qoraqalpoq tilidagi guruhlar va sahifalar kuniga yuzlab yangi matnlarni e’lon qilmoqda. Bu esa axborotni samarali boshqarish, qidirish va tahlil qilish uchun avtomatlashtirilgan tizimlarni joriy etish zarurligini ko‘rsatadi.

Matnlarni avtomatik tasniflash – tabiiy tilni qayta ishlash (Natural Language Processing-NLP) sohasini asosiy vazifalaridan biri bo‘lib, u yangilik agregatorlari, kontentni moderatsiya qilish, axborot qidirish va tavsiya tizimlari kabi ko‘plab amaliy masalalarni yechishda qo‘llaniladi [2]. Biroq qoraqalpoq tili uchun bunday tizimlar yo‘qligi nafaqat texnologik rivojlanishni sekinlashtirmoqda, balki til resurslarini raqamli shaklda saqlash imkoniyatlarini ham cheklaydi.

Mazkur tadqiqotning dolzarbligi quyidagilar bilan belgilanadi:

Birinchidan, qoraqalpoq tili uchun tabiiy tilni qayta ishlash sohasida ilmiy bo‘shliq mavjud. Xalqaro miqyosda amalga oshirilayotgan NLP tadqiqotlari deyarli faqat yirik tillarga qaratilgan bo‘lib, million va undan kam aholi so‘zlashadigan tillar e’tiborsiz qolmoqda. Shuningdek, qoraqalpoq tilida yaratilayotgan yangiliklar hajmi oshib bormoqda, biroq ularni avtomatik tarzda turkumlash, tahlil qilish va boshqarish uchun maxsus vositalar yo‘q. Bu esa media tashkilotlari, davlat organlari va tadqiqotchilar uchun jiddiy muammo hisoblanadi.

Aksariyat NLP modellari katta hajmdagi annotatsiyalangan ma’lumotlarga muhtoj. Qoraqalpoq tili uchun bunday resurslar deyarli mavjud emas. Shuning uchun klassik mashinali o‘qitish algoritmlari kamroq ma’lumot talab qiladigan, biroq yetarlicha samarali bo‘lgan vaziyatda eng maqbul yechim bo‘lib qolmoqda.

Tadqiqotni asosiy maqsadi qoraqalpoq tilidagi yangilik matnlarini avtomatik ravishda turkumlash uchun optimal mashinali o‘qitish modelini yaratish va resurs-cheklangan tillar uchun amaliy yechimlar taklif etishdir. Bu tadqiqot natijasida qoraqalpoq tilidagi raqamli kontentni boshqarish va tahlil qilish imkoniyatlari

ISSN 3030-3907

kengayadi, hamda boshqa kichik tillar uchun ham amaliy qo‘llanma bo‘lib xizmat qiladi.

METODOLOGIYA

Mazkur tadqiqotda qoraqalpoq tilidagi yangilik matnlarini avtomatik turkumlash vazifasini yechish uchun klassik mashinali o‘qitish usullaridan foydalanilgan. Tadqiqot jarayoni to‘rtta asosiy bosqichdan iborat: (1) ma’lumotlarni to‘plash va korpus yaratish, (2) matnlarni tozalash va qayta ishlash, (3) belgilarni shakllantirish, (4) modellarni o‘qitish va baholash.

Ma’lumotlar to‘plash va korpus yaratish. Tadqiqot uchun qoraqalpoq tilida faoliyat yurituvchi jaslarnews.uz [3] va kknews.uz [4] kabi ikkita asosiy internet resursi tanlangan. Ushbu saytlar yordamida yangilik maqolalari avtomatik ravishda yig‘ib olingan.

Matnlarni tozalash va dastlabki ishlov berish. Xom matnlarni qayta ishlash bir necha bosqichda amalga oshirilgan. Avval HTML teglari, URL manzillar va maxsus belgilar olib tashlanib, apostroflarni turli xil variantlari (’, ', `, ´) standart shaklga keltirilgan. Matnlar kichik harflarga o‘tkazilgan (lowercase). So‘ngra qoraqalpoq tili uchun maxsus to‘xtash so‘zlari (stop words) lug‘ati yaratilgan va bu so‘zlar matnlardan olib tashlangan.

Belgilarni shakllantirish. Tasniflash uchun uch xil belgilar to‘plami yaratilgan:

So‘z darajasidagi belgilar (Word-level features): TF-IDF vektorlashtirish usuli qo‘llanilgan. N-gramm diapazoni esa unigram, bigram va trigramlarni qamrab olgan.

Simvol darajasidagi belgilar (Character-level features): Belgi n-grammlari (3-5 belgi uzunlikdagi) TF-IDF usuli bilan vektorlashtirilgan, bu morfologik belgilarni ushlash imkonini beradi.

Lingvistik belgilar: So‘zlar soni, belgilar soni, o‘rtacha so‘z uzunligi, Type-Token Ratio (TTR) – leksik rang-baranglik ko‘rsatkichi, raqamlar soni, defis belgisi foizi kabi statistik ko‘rsatkichlar hisoblab chiqilgan.

Mashinali o‘qitish modellari. Tasniflash vazifasini yechish uchun klassik mashinali o‘qitishni beshta algoritmi tanlab olingan va taqqoslangan [5]:

1. Linear Support Vector Machine (Linear SVM) – chiziqli klassifikator, yuqori o‘lchamli fazolarda samarali ishlaydi.
2. Logistik regressiya – ehtimollik asosidagi klassifikator, maksimal 2000 iteratsiya bilan o‘qitildi.
3. Multinomial Naive Bayes (MNB) – Bayes teoremasiga asoslangan, matnlarni tasniflashda keng qo‘llaniladigan usul.
4. Random Forest – 300 ta qaror daraxti ansambli, parallel hisoblash qo‘llanildi.

ISSN 3030-3907

5. K-Nearest Neighbors (KNN) – k=7 yaqin qo‘shniga asoslangan klassifikator.

Baholash mezonlari. Modellar samaradorligi Accuracy, Precision, Recall, F1-score kabi metrikalar asosida baholangan.

NATIJALAR

Mazkur bo‘limda tadqiqot davomida olingan tajribalar natijalari taqdim etilgan bo‘lib, bunda jami 142 ta yangilik maqolasi to‘plangan. Korpusni kategoriyalar bo‘yicha taqsimoti quyidagi jadvalda keltirilgan.

1-jadval.

Korpusning kategoriyalar bo‘yicha taqsimoti

Kategoriya	Maqolalar soni	Ulushi (%)
Jamiyat	49	34.5
Siyosat	48	33.8
Iqtisod	18	12.7
Boshqa	17	12.0
Sport	6	4.2
Madaniyat	4	2.8
Jami	142	100.0

Mashinali o‘qitishning 5 ta modeli va ularning ansambl kombinatsiyasi validatsiya va test to‘plamlarida baholandi. Natijalar quyidagi jadvalda keltirilgan.

2-jadval.

Modellarning validatsiya va test to‘plamlaridagi natijalari

Model	Val Accuracy	Val F1-score	Test Accuracy	Test F1-score
Linear SVM	0.7273	0.6891	0.7143	0.6825
Logistic Regression	0.7727	0.7456	0.7619	0.7321
Multinomial NB	0.6364	0.5892	0.6190	0.5743
Random Forest	0.8182	0.7983	0.8095	0.7856
K-Nearest Neighbors	0.7273	0.6954	0.7143	0.6892
Ansambl	0.8182	0.7921	0.8095	0.7803

Natijalardan ko‘rinib turibdiki, eng yuqori ko‘rsatkichga Random Forest modeli erishdi: test to‘plamida accuracy=0.8095 va F1-score=0.7856. Ushbu modelni kategoriyalar kesimida ko‘rsatkichlari 3-jadvalda keltirilgan.

3-jadval.

Random Forest modelining kategoriyalar kesimida ko‘rsatkichlari

Kategoriya	Precision	Recall	F1-score	Namunalar soni
Jamiyat	0.88	0.78	0.83	9

ISSN 3030-3907

Siyosat	0.86	0.86	0.86	7
Iqtisod	0.67	1.00	0.80	2
Sport	1.00	0.50	0.67	2
Boshqa	0.00	0.00	0.00	1

Eng yuqori natija “Siyosat” kategoriyasida qayd etildi, “Jamiyat” kategoriyasi ham yaxshi tasniflandi. “Iqtisod” kategoriyasi yuqori recall (1.00) ko‘rsatdi, biroq precision past bo‘ldi. “Sport” kategoriyasida precision ideal (1.00), lekin recall past (0.50). “Boshqa” kategoriyasi uchun model yetarli o‘quv namunalari yo‘qligi sababli klassifikatsiya qila olmadi.

XULOSA

Qoraqalpoq tilidagi yangilik matnlarini mashinali o‘qitish algoritmlari yordamida tasniflash doirasida olib borilgan tadqiqot natijalaridan quyidagi xulosalar shakllandi:

– qoraqalpoq tilidagi yangiliklar korpusi turli xil internet manbalaridan to‘plandi va ular 6 ta kategoriyaga ajratildi;

– Linear SVM, Logistic Regression, Multinomial Naive Bayes, Random Forest, K-Nearest Neighbors kabi klassik mashinali o‘qitish algoritmlari va ularni ansambl kombinatsiyasi sinovdan o‘tkazildi;

– Random Forest modeli test to‘plamida accuracy=0.8095 va F1-score=0.7856 qiymati bilan eng yaxshi natijani ko‘rsatdi;

– kategoriyalar bo‘yicha tahlil shuni ko‘rsatdiki, “Siyosat” (F1=0.86) va “Jamiyat” (F1=0.83) kategoriyalari eng yaxshi tasniflash natijalarini ko‘rsatdi, “Sport” va “Iqtisod” kategoriyalarida esa o‘quv namunalarini hajmi kam bo‘lganligi uchun tasniflashda yaxshi natija ko‘rsatmadi.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI

1. N. A. Niyozmatova, N. S. Mamatov, B. I. Otaxonova, A. N. Samijonov and K. K. Erejepov, "Classification Based On Decision Trees And Neural Networks," 2021 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2021, pp. 01-04, doi: 10.1109/ICISCT52966.2021.9670345.

2. Nielsen, M. A., & Chuang, I. L. (2010). Quantum Computation and Quantum Information (10th Anniversary ed.). Cambridge University Press.

3. https://jaslarnews.uz/category/

4. https://kknews.uz/qq/category/news/

5. D.Muhamediyeva, N.Niyozmatova, N.Turgunova, A.Mamatov. Text vectorization based on the quantum bag-of-words algorithm. Electronic journal of actual problems of modern science, education and training. April, 2025-4. ISSN 2181-9750. Pp. 33-39.

ISSN 3030-3907