O‘ZBEKISTON MILLIY UNIVERSITETI XABARLARI, 2025, [1/11/1]
ISSN 2181-7324

FILOLOGIYA
http://journals.nuu.uz
Social sciences

UDK: 811.512.133’28:81’322

***Shahnoza POZILOVA,***
*Toshkent axborot texnologiyalari universiteti professori, DSc*
*E-mail:informatikpozilova@gmail.com*
***Madina RAXIMOVA,***
*Toshkent axborot texnologiyalari universiteti magistranti*
*E-mail:madinadilshodovna313@gmail.com*

*TATU dotsenti D.Zaripova taqrizi asosida*

## COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR IDENTIFYING UZBEK DIALECTS AT THE SENTENCE LEVEL

Annotation

This study examines the task of automatic classification of Uzbek language dialects. While resources of Natural Language Processing (NLP) are increasing, the scarcity of dialectological corpora remains one of the primary challenges. In this work, two fundamental approaches were tested on a small-scale, author-collected dataset comprising dialects on TF-IDF + Naive Bayes and BERT(bert-base-multilingual-cased) models. The main conclusion of the research is that the primary obstacle to creating high-accuracy models in Uzbek dialectology is not only the the right algorithm, but rather the absence of a high-quality, comprehensively annotated corpus.

**Key words:** Uzbek dialects, dialect classification, natural language processing, TF-IDF, Naive Bayes, BERT, mBERT, data scarcity.

## СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ИДЕНТИФИКАЦИИ УЗБЕКСКИХ ДИАЛЕКТОВ НА УРОВНЕ ПРЕДЛОЖЕНИ

Аннотация

Данное исследование посвящено задаче автоматической классификации диалектов узбекского языка. Несмотря на увеличение ресурсов в области обработки естественного языка (NLP), нехватка диалектологических корпусов остается одной из основных проблем. В настоящей работе на небольшом, собранном автором наборе данных, включающем диалекты, были протестированы два фундаментальных подхода: модели TF-IDF + Naive Bayes и BERT (bert-base-multilingual-cased). Главный вывод исследования заключается в том, что основным препятствием для создания высокоточных моделей в узбекской диалектологии является не столько выбор верного алгоритма, сколько отсутствие качественного, всесторонне аннотированного корпуса.

**Ключевые слова**: Диалекты узбекского языка, классификация диалектов, обработка естественного языка, TF-IDF, Naive Bayes, BERT, mBERT, нехватка данных.

## O‘ZBEK TILIDAGI SHEVALARNI GAP DARAJASIDA ANIQLASHDA MASHINALI O‘QITISH ALGORITMLARINING QIYOSIY TAHLILI

Annotatsiya

Ushbu tadqiqotda o‘zbek tili shevalarini avtomatik tasniflash vazifasi ko‘rib chiqiladi. Tabiiy tilni qayta ishlash (NLP) sohasida hozirda o‘zbek tili uchun resurslar kundan-kunga rivojlanib bormoqda va bu o‘z navbatida dialektologik korpuslarning yetishmasligi kabi muammolarni yuzaga keltiradi. Mazkur ishda kichik hajmdagi, turli shevalardan iborat ma’lumotlar to‘plami “Bag-of-Words” usuli bo‘lgan TF-IDF + Naive Bayes va bert-base-multilingual-cased modeli asosida tekshirildi. Tadqiqotning asosiy xulosasi, o‘zbek dialektologiyasida yuqori aniqlikdagi model qurilishidagi asosiy muammo bu faqatgina algoritm emas, balki yuqori sifatli, keng qamrovli annotatsiya qilingan korpusning yo‘qligidir.

**Kalit so‘zlar:** O‘zbek tili shevalari, dialektlarni tasniflash, tabiiy tilni qayta ishlash, TF-IDF, Naive Bayes, BERT, mBERT, ma’lumotlar yetishmasligi.

**Kirish.** O‘zbek tili dialektal xilma-xillikdan iborat, har bir hudud shevasi nafaqat fonetik, balki leksik va grammatik jihatlari bilan o‘zaro farqlanadi. Bunda shevalarni avtomatik aniqlash (Dialect Identification) raqamlashtirish jarayoni uchun dolzarb vazifalardan biriga aylanmoqda. So‘nggi yillarda o‘zbek tili uchun NLP sohasida sezilarli yutuqlarga erishildi, jumladan, UzBERT [1] kabi maxsus transformer modellari yaratildi. Biroq, mavjud tadqiqotlarning asosiy qismi adabiy tilga qaratilganligi sababli, sheva bilan bog‘liq vaziyatlarni tasniflashda biroz qiyinchiliklar mavjud. Ushbu ishning maqsadi o‘zbek tili shevalarini tasniflashda ikki xil avlod modellari ya’ni an’anaviy statistik model (TF-IDF) va zamonaviy transformer modelini (BERT) kichik ma’lumotlar to‘plami asosida taqqoshdan iborat.

**Tadqiqot usullari.** Tadqiqot uchun maxsus ma’lumotlar to‘plami yaratildi. Korpus 5 ta dialekt (Adabiy, Toshkent, Farg‘ona, Samarqand, Xorazm), jami ~120-130 ta gapdan iborat. TF-IDF + Naive Bayes klassik yondashuv TF-IDF (Term Frequency-Inverse Document Frequency) matnlar bilan ishlashda so‘zlarning chastotasiga asoslangan vektorizatoridan o‘tkazildi. So‘ngra, ushbu vektorlar Multinomial Naive Bayes (MNB) klassifikatoriga o‘qitildi. Kerakli kutubxonalar yuklanib, ma’lumotlar tozalanadi va

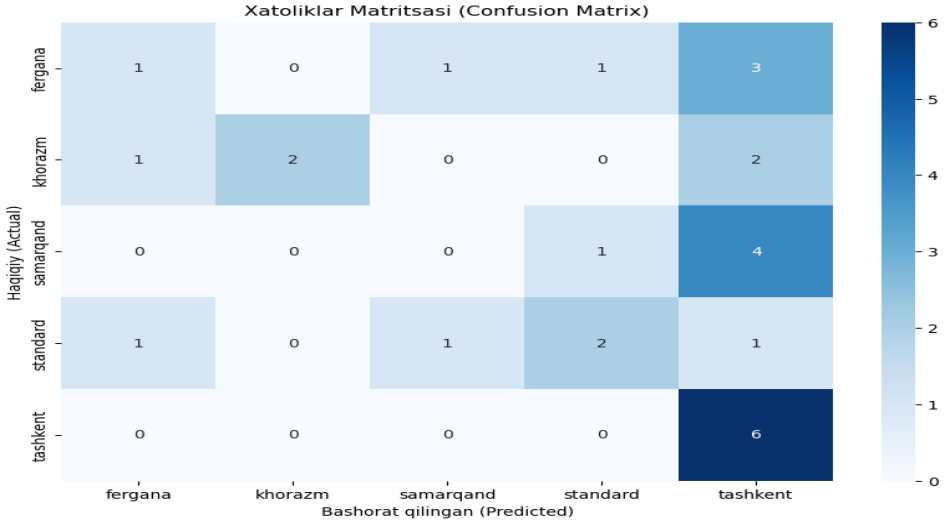
tadqiqot uchun “test” va “train” qismlar TF-IDF modeilda o‘qitildi. Quyida dasturning asosiy qismi keltiriladi:

Dastur kodi: (Python dasturlash tilida, Jupyter Notebook)

```
text_clf_pipeline = Pipeline([('tfidf', TfidfVectorizer(ngram_range=(1, 2))), ('clf', MultinomialNB()),# 2. Naive Bayes klassifikatorini qo‘llash ])
text_clf_pipeline.fit(X_train, y_train)
print("Model muvaffaqiyatli o‘qitildi.")
print("-" * 30)
Natija: Model muvaffaqiyatli o‘qitildi.
y_pred = text_clf_pipeline.predict(X_test)
print("Boshlang‘ich Model Natijalari (TF-IDF + Naive Bayes):")
print(classification_report(y_test, y_pred))print("-" * 30)
```

Natija: Boshlang‘ich Model Natijalari (TF-IDF + Naive Bayes):

```
precision recall f1-score support
fergana 0.33 0.17 0.22 6
khorazm 1.00 0.40 0.57 5
samarqand 0.00 0.00 0.00 5
standard 0.50 0.40 0.44 5
tashkent 0.38 1.00 0.55 6
accuracy 0.41 27
macro avg 0.44 0.39 0.36 27
weighted avg 0.44 0.41 0.36 27
cm = confusion_matrix(y_test, y_pred)
class_names = sorted(y.unique())
plt.figure(figsize=(10, 7))
sns.heatmap(cm, annot=True, fmt='g‘, cmap='Blues', xticklabels=class_names, yticklabels=class_names)
plt.xlabel('Bashorat qilingan (Predicted)')
plt.ylabel('Haqiqiy (Actual)')
plt.title('Xatoliklar Matritsasi (Confusion Matrix)')
plt.show()
```




1-rasm: Xatoliklar matritsasi

Keyingi model uchun 104 ta tilni, shu jumladan o‘zbek tilini ham o‘z ichiga olgan bert-base-multilingual-cased (mBERT) [4] modelidan foydalanildi. Model 5 ta epoxa davomida kichik ma’lumotlar to‘plamiga qayta o‘qitildi. Kerakli kutubxonalar yuklanib, ma’lumotlar tozalandi va tadqiqot uchun “test” va “train” qismlar modelda o‘qitildi. Quyida dasturning asosiy qismi keltiriladi:

Dastur kodi: (Python dasturlash tilida, GOOGLE Collab)

```
model_name = "bert-base-multilingual-cased"
print(f"\nModel yuklab olinmoqda: {model_name}")
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForSequenceClassification.from_pretrained( model_name, num_labels=len(label_names) # Bizda 5 ta sheva bor (fergana, khorazm, ...) )
print("Model muvaffaqiyatli yuklandi.")
def tokenize_function(examples):
return tokenizer(examples['text'], padding='max_length', truncation=True)
tokenized_datasets = dataset_dict.map(tokenize_function, batched=True)
print("\nMa’lumotlar tokenizatsiya qilindi.")
def compute_metrics(eval_pred): logits, labels = eval_pred
predictions = np.argmax(logits, axis=-1)
precision, recall, f1, _ = precision_recall_fscore_support(labels, predictions, average='weighted', zero_division=0)
acc = accuracy_score(labels, predictions)
return { 'accuracy': acc, 'f1': f1,'precision': precision, 'recall': recall }
training_args = TrainingArguments( output_dir="./results",eval_strategy="epoch",save_strategy="epoch",learning_rate=2e-5,per_device_train_batch_size=8,per_device_eval_batch_size=8, num_train_epochs=5, weight_decay=0.01,load_best_model_at_end=True, metric_for_best_model="f1",report_to="none",
trainer = Trainer( model=model,args=training_args,train_dataset=tokenized_datasets["train"],
eval_dataset=tokenized_datasets["test"],compute_metrics=compute_metrics,)
print("\n--- MODELNI O‘QITISH BOSHLANDI ---")
trainer.train()
print("--- MODELNI O‘QITISH TUGADI ---")
Natija: --- MODELNI O‘QITISH BOSHLANDI ---
{'eval_loss': 1.5220143795013428, 'eval_accuracy': 0.3333333333333333, 'eval_f1': 0.2160493827160494, …
--- MODELNI O‘QITISH TUGADI ---
[ ] print("\nEng yaxshi modelning test ma’lumotlaridagi yakuniy natijalari:")
eval_results = trainer.evaluate() print(eval_results)
```

Natija: Eng yaxshi modelning test ma’lumotlaridagi yakuniy natijalari:

{'eval_loss': 1.0687419176101685, 'eval_accuracy': 0.5185185185185185, 'eval_f1': 0.4423497230514774, 'eval_precision': 0.5197293447293447, 'eval_recall': 0.5185185185185185, 'eval_runtime': 0.8201, 'eval_samples_per_second': 32.923, 'eval_steps_per_second': 4.878, 'epoch': 5.0}

{'eval_loss': 1.0687419176101685, 'eval_accuracy': 0.5185185185185185, 'eval_f1': 0.4423497230514774, 'eval_precision': 0.5197293447293447, 'eval_recall': 0.5185185185185185, 'eval_runtime': 0.8201, 'eval_samples_per_second': 32.923, 'eval_steps_per_second': 4.878, 'epoch': 5.0}

print(f"\nYakuniy aniqlik (Accuracy): {eval_results['eval_accuracy'] * 100:.2f}%")

print(f"Yakuniy F1-score (Weighted): {eval_results['eval_f1'] * 100:.2f}%")

Natija: Yakuniy aniqlik (Accuracy): 51.85% Yakuniy F1-score (Weighted): 44.23%

NATIJALAR TAHLILI

TF-IDF modeli “qivossan” (Toshkent) va “qilyapsan” (Adabiy) so‘zlarini ikki xil, mutlaqo bog‘liq bo‘lmagan birlik sifatida ko‘radi. U “Ukam galdi” (Xorazm) va “Ukam keldi” (Adabiy) gaplari orasidagi kontekstual yaqinlikni tushunmaydi. Ma’limotlar to‘plamida(kichik), agar “test” qismida “individual” kalit so‘zlar (gun, yatbti) tushib qolsa, u yaxshi natija berishi mumkin (“Xorazm” deb bashorat qilgan barcha gaplar aynan Xorazm shevasi bo‘lgan. Aksincha bir-biriga o‘xshash Qarluq shevalari (Toshkent, Farg‘ona, Samarqand) tushib qolsa, to‘liq muvaffaqiyatsizlikka uchraydi. Bu shuni anglatadiki, TF-IDF + Naive Bayes yondashuvi gap darajasidagi ushbu tahlil uchun ishonchsizdir.

Keyingi transformer model BERT orqali 51.85% aniqlikka erishildi va quyidagi ikki muhim cheklov aniqlandi: Model - bu yerda UzBERT emas, balki mBERT (ko‘p tilli) modeli ishlatildi, mBERT 104 ta tilni biladi, lekin u o‘zbek tilining nozik dialektal farqlari bo‘yicha yetarlicha samarali emas. Ma’lumotlar - eng asosiy muammo ma’lumotlar hajmida bo‘ldi. Tadqiqotda 177 million parametrli ulkan arxitekturani taxminan 100 ta gapdan iborat o‘qitish to‘plamiga o‘rgatishga harakat qilindi. Bu, modelning shevalar orasidagi nozik farqlarni o‘rganishi uchun mutlaqo yetarli emasligini ko‘rsatadi. Model shunchaki “past darajada o‘qitilgan” (underfitted) bo‘lib qolmoqda.

**Xulosa.** Ushbu tadqiqotdan shunday xulosa shuki, o‘zbek shevalarini avtomatik tasniflashdagi asosiy to‘siq algoritmning (modelning) kuchsizligi emas, balki yuqori sifatli, keng qamrovli va annotatsiya qilingan dialektal korpusning yo‘qligidir. Shuni ham ta’kidlab o‘tish joizki tilning nozik tuzilishi va shevalar xilma-xilligini inobatga olib, ushbu modellar so‘z darajasida emas balki gap darajasida qiyinchiliklarga uchramoqda. Bunda sifatni yana ham yaxshilash va aniq natijalar olishda nafaqat juda katta ma’lumotlar bazasi balki nozik til qoidalariga mos ravishda ishlay oladigan UzBERT kabi maxsus o‘rgatilgan shevalarni 90%+ barqaror aniqlikda tasniflay olish salohiyatiga ega bo‘lgan modellarni yanada takomillashtirish maqsadga muvofiq.

## ADABIYOTLAR


1. M. K., S. A., T. O., & K. M. (2021). UzBERT: A New Uzbek Language Model and Its Application in Sentiment Analysis.
2. Mansurov B.., A. Mansurov. (2021). UzBERT: pretraining a BERT model for Uzbek. Copper City Labs
3. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. Article in Journal of Machine Learning Research
4. Wolf, T. et al. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Hugging Face, Brooklyn, USA
5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics.
6. Reshetov V.V., Sh.Shoabdurahmonov muallifligida yaratilgan “O‘zbek dialektologiyasi” darsligining 60 yilligi munosabati bilan tashkil etilgan. (2022) “O‘zbek shevalari tadqiqotlari: amaliyot, metodologiya va yangicha yondashuv” mavzusidagi II Respublika ilmiy-nazariy konferensiyasi materiallari. Toshkent «Donishmand ziyosi»
7. Abdulla Qahhor. (1936). Anor, O‘g‘ri, Bemor, Daxshat hikoyalari. (adabiy sheva uchun)
8. Nazar Eshonqul. (2004).Maymun yetaklagan odam hikoyasi. (adabiy sheva uchun)
9. Ijtimoiy tarmoqlar: Instagram, Telegram (mahalliy aholining shevasi)