

**Matnli hujjatlarni Mashinali o'rgatish usullari yordamida tasniflash****Saidov Doniyor****Yusupov Jaloliddin**

Aniq va ijtimoiy fanlar universiteti. Toshkent, O'zbekiston

doniyorsaidov86@gmail.comyusupovjaloliddin4633@gmail.com

Anotatsiya: Mazkur tadqiqot ishida matnli hujjatlarni tasniflash uchun mashinali o'rgatish usullari – Logistic Regression, Naive Bayes va Long short-term memory usullari tahlil qilindi. Hisoblash tajribalarida matematika, fizika, huquqshunoslik va biologiya fanlariga tegishli hujjatlar sinovdan o'tkazildi va natijalar olindi.

Kalit so'zlar: Matnli hujjatlarni tasniflash, Logistic Regression (LR), Naive Bayes (NB), Long short-term memory (LSTM):, mashinali o'rgatish, tabiiy tilni qayta ishlash.

Tabiiy tilni qayta ishlash va matnlarni intellektual tahlil qilish sohasida turli xil mashinali o'rgatish usullari yordamida matnli hujjatlarni tasniflash eng ko'p o'rganiladigan masalalardan biri hisoblanadi. Odatda dastlabki berilganlar sifatida turli sohalarga oid hujjatlarning chekli to'plami beriladi. Ushbu tadqiqot ishida barcha hujjatlar .txt yoki .doc,.docx kengaytmali fayllar yordamida berilgan va tanlanma obyektlari shakllantirilgan. O'rgatuvchi tanlanmada har bir hujjat bitta obyektini ifodalaydi, soha esa sinf sifatida olingan.

Mashinali o'rganish usullari yordamida hujjatlarni avtomatik tasniflash nafaqat vaqtni tejaydi, balki tasniflash jarayonining aniqligini oshiradi. Ushbu tadqiqot ishida mashinali o'rgatishning quyidagi usullari o'rganildi hamda chuqur tahlil qilindi:



Logistic Regression (LR): chiziqli model asosida ishlaydigan tasniflash algoritmi bo'lib, kichik hajmdagi matnlarni tasniflash masalalarida tez-tez qo'llaniladi.

Naive Bayes (NB): ehtimollik nazariyasiga asoslangan algoritm bo'lib, hujjatlarni tasniflashda sodda va samarali usul hisoblanib xususiyatlar orasidagi bog'liqlikni e'tiborga olmagan holda tasniflashni amalga oshiradi.

Long short-term memory (LSTM): chuqur o'rganish (deep learning) algoritmlari oilasiga kiruvchi, rekurrent neyron tarmoqlarning (RNN) maxsus turi hisoblanadi. Bu usul ayniqsa matnlar va ketma-ketliklarga ishlov berishda samarali.

Hisoblash eksperimenti sifatida oliy attestatsiya web saytidan olingan matematika, fizika, huquqshunoslik va biologiya fan tarmoqlariga tegishli avtoreferatlar jamlanmasi qaraldi. Eksperiment ikki xil tanlanma asosida o'tkazildi. Birinchi eksperimentda har bir sohadan, ya'ni matematikaga doir 20 ta, fizikaga doir 20 ta, huquqshunoslikka doir 20 ta va biologiyaga doir 20 tadan fayllar tanlanmani shakllantirish uchun olindi. Ikkinchi eksperimentda mos ravishdan 3 ta sohadan ya'ni matematikadan 64 ta, huquqshunoslikka doir 84 ta va biologiyadan 107 ta xujjatlar asosida tanlanma shakllantirildi. O'rgatuvchi tanlanma sifatida jami obyektlarning 80% tasodifiy ravishda (one-hold-out asosida) ajratib olindi va qongan 20% qismi nazorat tanlanmasi sifatida olindi.

Jadval №1. Algoritmilar sifatini baholash.

T/r	Usul	Accuracy	Recall	Precision	F1-score
1	LR	0.94	0.92	0.92	0.92
2	NB	0.83	0.89	0.83	0.82
3	LSTM (1-tanlanma)	0.75	0.69	0.69	0.69
4	LSTM (2-tanlanma)	0.90	0.90	0.90	0.90

Yuqoridagi natijalardan ko'rish mumkinki (1-jadval) LR usuli NB usulidan ko'ra barcha mezonlar bo'yicha sezilarli yaxshi natija ko'rsatdi. LSTM usulida esa katta xajmli tanlanma kichik hajmli tanlanmaga nisbatan yaxshi natija qayt qildi.



Foydalanilgan adabiyotlar:

1. N Ignatev, D Saidov, U Tuliev. [Cluster analysis of document collections based on topological properties of objects](#) // AIP Conference Proceedings, 2024, Volume 3244, Issue 1. <https://doi.org/10.1063/5.0242564>
2. D Saidov, S Eshbadalov. Analysis of classification of text documents according to the content using machine learning methods: Analysis of classification of text documents according to the content ... MODERN PROBLEMS AND PROSPECTS OF APPLIED MATHEMATICS, 2024/6/7, Volume 1, Issue 01.