ISSN 2181-3187

MASHINALI O'RGATISH USULLARI YORDAMIDA MATNLI HUJJATLARNI MAZMUNIGA KO'RA TASNIFLASH

Eshbadalov Sirojiddin Usmon o'g'li

Toshkent Kimo Xalqaro Universiteti

"Sun'iy intellekt" mutaxassisligi magistratura talabasi

Email: eshbadalovsirojiddin@gmail.com

Annotatsiya: Mazkur tadqiqot ishida matnli hujjatlarni mazmuniga ko'ra avtomatik tasniflash uchun mashinali o'rgatish usullaridan biri – O'rama neyron tarmoqlari (CNN) tahlil qilindi. Tadqiqot jarayonida yangiliklar saytlaridan olingan turli mavzularga oid matnli maqolalardan tashkil topgan datasetdan foydalanildi. CNN algoritmi yordamida ushbu ma'lumotlar ustida amaliy tajribalar o'tkazildi. Olingan natijalar asosida CNN usulining samaradorligi klassik mashinali o'rgatish algoritmlari bilan taqqoslanib, mazkur yondashuvning matnlarni avtomatik tasniflash vazifasida yuqori aniqlik va samaradorlik kabi afzalliklari mavjudligi ko'rsatib berildi. Natijalar CNN modelining matnli hujjatlarni tasniflash sohasidagi istiqbolli ekanligini isbotladi.

Kalit so'zlar: Matnlarni tasniflash, mashinali o'rgatish, O'rama neyron tarmoqlari (CNN), CNN texnikalari, konvolyutsiya, pooling, embedding, yangiliklar maqolalari, avtomatik tasniflash, algoritmlar taqqoslanishi, aniqlik, samaradorlik.

Mashinali o'rganish usullari yordamida matnli hujjatlarni avtomatik tasniflash nafaqat vaqtni tejaydi, balki tasniflash jarayonining aniqligini ham sezilarli darajada oshiradi. Ushbu tadqiqot ishida mashinali o'rgatishning quyidagi usuli o'rganildi hamda chuqur tahlil qilindi: O'rama neyron tarmoqlari (CNN). Tadqiqot davomida yangiliklar saytlari orqali olingan turli mavzulardagi matnli maqolalar ustida CNN algoritmidan foydalangan holda amaliy tajribalar o'tkazildi. Olingan natijalar boshqa

ISSN 2181-3187

klassik algoritmlar bilan solishtirilib, CNN usulining samaradorligi va yuqori aniqligi namoyon etildi.

CNN (Convolutional neural network) - usuli matnli hujjatlarni mazmuniga ko‘ra tasniflashda quyidagi tarzda amalga oshiriladi: dastlab matnli ma’lumotlar embedding qatlami yordamida vektor ko‘rinishiga keltiriladi, so‘ngra konvolyutsiya (convolution) qatlamlari orqali matn tarkibidagi xususiyatlar (features) aniqlanadi va pooling qatlamlari orqali muhim belgilar saralanadi. Nihoyat, to‘liq bog‘langan (fully-connected) neyron qatlamlari orqali matnlar sinflarga ajratilib, avtomatik tasniflash amalga oshiriladi. Tadqiqot davomida yangiliklar saytidan olingan matnli maqolalar CNN algoritmi orqali sinovdan o‘tkazilib, uning samaradorligi va aniqligi ko‘rsatib berildi.

Tadqiqotda umumiy 3200 ta matnli hujjatdan iborat datasetdan foydalanildi. Ushbu hujjatlar yangiliklar saytlaridan olingan bo‘lib, ularning har biri aniq bir mavzuga oid bo‘lib, jami 7 ta sinf (kategoriya)ga tegishli. Matnli hujjatlarni mazmuniga ko‘ra tasniflash vazifasini amalga oshirish uchun hisoblash eksperimenti quyidagi asosiy bosqichlarda olib borildi:

1. Ma’lumotlarni oldindan qayta ishlash (Preprocessing):

- Tozalash (Cleaning): Har bir matndan tinish belgilar, raqamlar, maxsus belgilar va HTML teglar olib tashlandi.
- Kichik harflarga o‘tkazish: Barcha matnlar kichik harflarga o‘tkazildi (lowercasing) – bu matnni normalizatsiyalashga xizmat qildi.
- Tokenlash: Matnlar so‘z birliklariga ajratildi (tokenization).
- Stop-so‘zlarni olib tashlash: Masalan ingliz tilidagi umumiy ma’no bermaydigan stop-so‘zlar (e.g., "the", "and", "is") olib tashlandi. O’zbek tilidagi ahamiyatsiz so’zlar (stop –words), ular (“va”, “lekin”, “biroq” va hokazo).
- Lemmatizatsiya: Har bir token o‘zining asosiy (lug‘aviy) shakliga keltirildi.

ISSN 2181-3187

- Label encoding: Kategoriyalar raqamli formatga o‘tkazildi (0 dan 6 gacha).

2. Vektorlash (Embedding):

- So‘zlarni raqamli ko‘rinishga o‘tkazish uchun Keras Embedding Layer ishlatildi.
- Alternativ tarzda, pre-trained GloVe embedding (Global Vectors for Word Representation) modelidan ham foydalanildi, bu model matnlardagi semantik aloqalarni yaxshiroq ushlashga yordam berdi.

3. Model arxitekturasi (CNN Modeling):

- Embedding layer: Matnlar 100–200 o‘lchamli vektorlar orqali ifodalandi.
- Convolutional layer: 1D konvolyutsiya qatlamlari (e.g., Conv1D) orqali n-gram (so‘z guruhlari) xususiyatlari chiqarib olindi.
- Activation function: ReLU (Rectified Linear Unit) faollashtirish funksiyasi qo‘llanildi.
- MaxPooling layer: MaxPooling1D yordamida eng muhim xususiyatlar ajratib olindi va o‘lcham kamaytirildi.
- Dropout layer: 0.5 ehtimollikda Dropout qatlam qo‘llanib, ortiqcha moslashuv (overfitting) oldi olindi.
- Flatten layer: Konvolyutsion chiqish natijasi tekislashtirildi.
- Fully Connected (Dense) layer: Sinflarni aniqlash uchun oxirgi qatlamda Dense(7, activation='softmax') ishlatildi.

4. Modelni o‘qitish va baholash:

- Dataset 3 qismga bo‘lindi: 70% – o‘qitish (training), 15% – tasdiqlash (validation), 15% – test (sinov).
- Yo‘qotish funksiyasi sifatida categorical crossentropy, optimallashtirish uchun esa Adam optimizer tanlandi.
- Model 10–20 epoch davomida o‘qitildi, har bir epochda aniqlik (accuracy) va yo‘qotish (loss) baholandi.

ISSN 2181-3187

- Erta to‘xtatish (EarlyStopping) texnikasi orqali optimal natijalar kuzatildi.

5. Natijalarni baholash:

- Model test to‘plamida 7 sinfga nisbatan umumiy aniqlik (accuracy), aniqlik (precision), qayta chaqirish (recall) va F1-mezon bo‘yicha baholandi.
- CNN modeli, ayniqsa konvolyutsion qatlamlar orqali matndagi semantik va sintaktik naqshlarni samarali aniqlashi natijasida, klassik usullarga (masalan, Logistic Regression, Decision Tree, Naive Bayes) nisbatan yuqori aniqlik ko‘rsatdi.

1 – jadval. Algoritmlarning baholash ko‘rsatkichlari

T/r	Usul	Accuracy (%)	Recall	F1-score
1	Decision Tree	82	0.80	0.81
2	Naive Bayes	88	0.86	0.87
3	SVM	85	0.83	0.84
4	CNN	91	0.90	0.91

Yuqoridagi jadvalda keltirilgan natijalardan ko‘rinib turibdiki, to‘rt xil mashinali o‘rgatish algoritmi – **Decision Tree**, **Naive Bayes**, **SVM** va **CNN** – matnli hujjatlarni mazmuniga ko‘ra tasniflash vazifasida sinovdan o‘tkazildi. Baholash mezonlari sifatida **Accuracy (%), Recall** va **F1-score** ko‘rsatkichlari asos qilib olindi.

Tajriba natijalariga ko‘ra, **CNN (Convolutional Neural Network)** modeli eng yuqori natijalarni ko‘rsatdi:

- **Accuracy:** 91%
- **Recall:** 0.90
- **F1-score:** 0.91

ISSN 2181-3187

Bu ko‘rsatkichlar CNN modelining matndagi murakkab semantik va sintaktik naqshlarni chuqur aniqlash qobiliyatiga ega ekanligini ko‘rsatadi.

CNN modeli quyidagi **asosiy afzalliklari** bilan ajralib turadi:

1. **Konvolyutsion qatlamlar** orqali matndagi n-gramlar va kontekstual aloqalarni avtomatik aniqlaydi, bu esa klassik algoritmlardan ancha samaraliroq bo‘ladi.
2. **Feature extraction (xususiyatlarni ajratish)** jarayoni modelning o‘zi tomonidan o‘rganiladi, foydalanuvchidan qo‘lda tanlash talab qilinmaydi.
3. **Overfitting (ortiqcha moslashish)** ni kamaytirish uchun Dropout va Pooling kabi qatlamlar samarali qo‘llaniladi.
4. **Parallel computation** imkoniyatlari tufayli katta datasetlarda ham yaxshi natijalar beradi (GPU bilan ishlaganda ayniqsa).
5. Matnli hujjatlar ustida ishlashda **semantik aniqlikni** yuqori darajada ta'minlaydi.

Shu asosda, matnli hujjatlarni mazmuniga ko‘ra avtomatik tasniflashda CNN modeli **eng istiqbolli va samarali yondashuv** sifatida tanlandi.

2 - jadval. Mashinali o‘rgatish algoritmlarining trening murakkabligi va bajarilish samaradorligi taqqoslanishi

/r	Model	Yondashuv turi	O‘rganiladigan parametrlar	Training usuli	Vaqt sarfi
	Naive Bayes	Statistik	Juda kam	Zudlik bilan (instant)	Juda tez
	Decision Tree	Qoidaviy	Past darajadagi qoidalar	Recursive splitting	Tez

ISSN 2181-3187

	SVM	Yuzaga asoslangan	O‘rta	Iterativ	O‘rtacha
	CNN	Neyron tarmoq	Juda ko‘p	Gradient descent	Uzoq

Tadqiqotda matnli hujjatlarni mazmuniga ko‘ra tasniflash uchun Decision Tree, Naive Bayes, SVM va CNN algoritmlari tahlil qilindi. 3200 ta yangilik maqolalari asosida o‘tkazilgan tajriba natijalariga ko‘ra, CNN modeli eng yuqori aniqlik (91%) va F1-score (0.91) ko‘rsatkichlariga erishdi. CNN matndagi semantik bog‘liqliklarni chuqur o‘rganishi bilan ajralib turadi. Shuning uchun CNN mazmuniy tasniflashda eng samarali va istiqbolli yondashuv sifatida tanlandi.

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.

3. Kim, Y. (2014). *Convolutional Neural Networks for Sentence Classification*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746–1751. https://doi.org/10.3115/v1/D14-1181.

4. Zhang, X., Zhao, J., & LeCun, Y. (2015). *Character-level Convolutional Networks for Text Classification*. In Advances in Neural Information Processing Systems (NeurIPS), 28, 649–657.

5. Matnli hujjatlarni Mashinali o‘rgatish usullari yordamida tasniflash. Saidov Doniyor, Yusupov Jaloliddin. Vol.1No12(2024). Dece