RAQAMLI TEXNOLOGIYALARNING NAZARIY VA AMALIY MASALALARI XALQARO JURNALI

P-ISSN: 2181-3086 E-ISSN: 2181-3094

Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti Samarqand filiali

Web: https://ijdt.uz/index.php/ijdt

SUN'IY INTELLEKT ALGORITMLARI ASOSIDA MATN TILINI AVTOMATIK ANIQLASH

Ilyos Xujayarov[1], Mannon Ochilov[2], Orzimurod Xolmatov[2], Dilshod Jurayev [2]

[1] Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti Samarqand filiali, Samarqand, O'zbekiston
[2] Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti, Toshkent, O'zbekiston
i.khujayorov@tuit.uz, ochilov.mannon@mail.ru, xolmatov.orzumurod@gmail.com, dilsamtuit@tuit.uz

Citation: *Xujayarov, I., Ochilov, M., Xolmatov, O., & Jurayev, D. (2024). Sun'iy intellekt algoritmlari asosida matn tilini avtomatik aniqlash. Международный Журнал Теоретических и Прикладных Вопросов Цифровых Технологий, 7(2), 59–67. https://doi.org/10.62132/ijdt.v7i2.182*

Kelib tushdi: 6-aprel 2024-yil
Qabul qilindi: 24-aprel 2024-yil
Chop etildi: 30-iyun 2024-yil

DOI: https://doi.org/10.62132/ijdt.v7i2.182

UDK 004.89

SUN'IY INTELLEKT ALGORITMLARI ASOSIDA MATN TILINI AVTOMATIK ANIQLASH

Xujayarov I.Sh.[1], *Ochilov M.M.*[2], *Xolmatov O.A.*[2], *Jurayev D.B.*[2]
[1] Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti Samarqand filiali, Samarqand, O'zbekiston
[2] Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti, Toshkent, O'zbekiston
i.khujayorov@tuit.uz, ochilov.mannon@mail.ru, xolmatov.orzumurod@gmail.com, dilsamtuit@tuit.uz

Annotatsiya. *Maqolada matn tilini aniqlashning sun'iy intellekt algoritmlariga asoslangan yondashuvlari muhokama qilinadi. Matn tilini aniqlash sun'iy intelletkning sinflashtirish masalasi bo'lganligi sababli, maqolada mashinali o'qitish va neyron tarmoq modellarining bir nechta sinflashtirish algoritmlari imkoniyatlari ko'rib o'tiladi. Ishda o'zbek, inlgiz, rus, qoraqalpoq tillarini aniqlovchi model uchun o'quv ma'lumotlar to'plamini shakllantirish masalasi ko'riladi. Shuningdek matn tilini aniqlashda foydalanilgan modellarning aniqlik ko'rsatkichlari bo'yicha qiyosiy tahlil amalga oshiriladi.*

Kalit so'zlar: *NLP, crowling, matn tili, o'quv ma'lumotlar to'plami, mashinali o'qitish, sinflashtirish, chuqur o'qitish, rekurrent neyron tarmoqlar, model aniqligini baholash.*

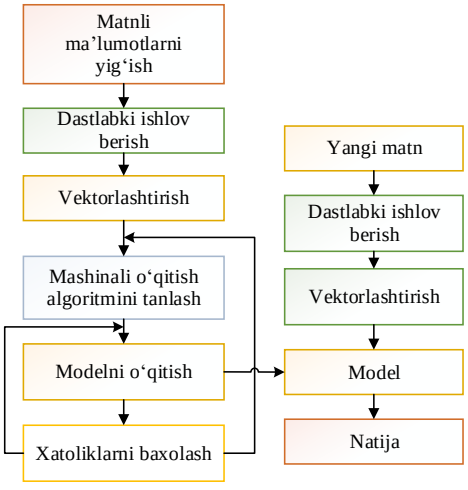
I. KIRISH

Bugungi taraqqiy etgan davrda insonlar bir birini tushunishi muhim ahamiyat kasb etadi. Yozma muloqotni amalga oshirish uchun esa matn qaysi tilda ekanligini aniqlash muhim masalalardan biri hisoblanadi [1]. Matn tilini aniqlash tabiiy tilni qayta ishlashda (NLP – Natural language processing) asosiy vazifalardan bo'lib, uni crawling (veb saytlardan matnlarni ajratib olish) jarayonlarida, mashina tarjimasi, savol - javob tizimlari, matnni tahlil qilish va ma'lumotlarni qidirish kabi turli sohalarda keng qo'llashmoqda. Garchan matn tilini aniqlash bo'yicha ko'plab tadqiqotlar amalga oshirilgan va yaxshi natijalarga erishilgan bo'lsada, ammo bu tadqiqotlarning asosiy qismi ingliz, rus, xitoy, fransuz, hindi kabi dunyo bo'ylab ko'p aholi foydalanadigan tillar ustidagi tadqiqotlarga asoslangan. NLP sohasida bu kabi tadqiqot ishlarini amalga oshirish uchun eng asosiy bo'g'inlardan biri bu - to'g'ri shakllantirilgan ma'lumotlar hisoblanadi.

Matnli ma'lumotlarni shakllantirishda asosan turli kitoblardan va veb saytlardan ma'lumotlar yig'iladi. Veb saytlardan faqat O'zbek tilidagi matnlarni ajratib olishda qiyinchiliklar yuzaga kelishi mumkin. Sababi ko'pgina veb saytlar bitta tilda emas, balki bir nechta tillarda faoliyat olib borish uchun mo'ljallangan bo'ladi. O'zbekiston-dagi veb saytlar asosan o'zbek, rus, ingliz va qoraqalpoq tillarida faoliyat olib boradi. Ulardan kerakli ma'lumotlarni olish uchun veb sahifadagi matnning qaysi tilda yozilganligini aniqlab olish masalasi turadi. Ko'pgina boshqa tillar kabi o'zbek tilida ham matnning qaysi tilda yozilganligini avtomatik aniqlash ishlari yetarlicha amalga oshirilmagan. Ushbu maqolada ingliz, rus, qoraqalpoq va o'zbek tilidagi matnlardan tilni avtomatik aniqlash jarayonlari, tilni aniqlash usullari, algoritmlari va taqdiqot mobaynida qo'lga kiritilgan natijalar keltirilgan.

II. ASOSIY QISM

NLP da tilni aniqlashning ko'plab mashinali o'qitish (MO') usullari mavud bo'lsada, lekin amalga oshirish jarayonlar ketma-ketligi umumiy hisoblanadi va uni quyidagicha tasvirlash mumkin:



1-rasm. (MO') algoritmlari yordamida tilni aniqlash sxemasi

Matnli ma’lumotlarni yig‘ish bosqichi eng murakkab bosqichlardan hisoblanib, bajarilishi kerak bo‘lgan loyihaga mos ravishda ma’lumotlarni yig‘ishni o‘z ichiga oladi. Bunda asosiy ma’lumotlar manbai bo‘lib, muayyan amaliy masalani (tilni aniqlashni) hal qilish uchun zarur bo‘lgan ma’lumotlarni o‘z ichiga olgan obyekt tushuniladi. Ma’lumotlarni to‘plash matn yoki hujjatlar shaklida katta hajmdagi ma’lumotlarni yig‘ish bilan boshlanadi. Bu ma’lumotlar kitoblar, internet saytlardan, maqolalar, tez-tez so‘raladigan savollar va boshqalar kabi turli manbalardan bo‘lishi mumkin [2].

Ma’lumotlarga dastlabki ishlov berish bosqichida yaratilayotgan tizimining aniqlilik darajasini oshirish maqsadida to‘plangan matnlar ichida matn ma’nosiga ta’sir qilmaydigan ma’lumotlar tozalanadi va oldindan qayta ishlanadi [3, 4]. Bular stop words (nomuhim so‘zlar), shovqin va nomuvofiqliklar (ba’zi bir belgilar qo‘shilib ketishi, so‘zlardagi xatoliklar) bo‘lishi mumkin [2]. Ushbu bosqichda matnli ma’lumotlar hammasi tozalanadi.

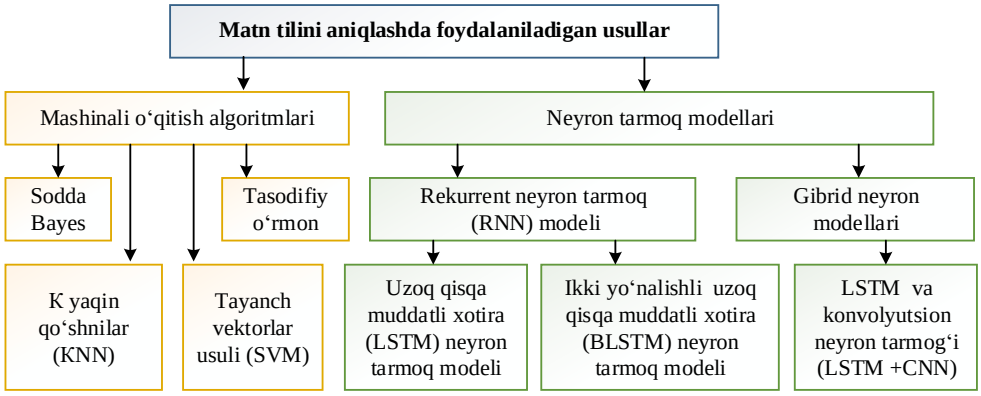
Ma’lumotlarni vektorlashtirish bosqichida N-gramm, atama chastotalari TF (Term frequency) yoki atama chastotasi - teskari hujjat chastotasi TF-IDF (Term frequency - inverse document frequency) kabi usullardan foydalangan holda matnni raqamli vektorlarga aylantirish kabi amallar bajariladi. Ushbu tadqiqotni amalga oshirish mobaynida kiruvchi matnli ma’lumotlarni TF-IDF vektoriga o‘tkazib olish amalga oshirildi. Olib borilgan tahlillar natijasida matn tilini aniqlashda ushbu xususiyatlar vektori yuqori samara berganini ko‘rish mumkin [5].

Mashinali o‘qitish algoritmini tanlash bosqichida masalani hal qilish uchun foydalaniladigan algoritm tanlanadi. Ishda sinflashtirish masalalari uchun yuqori samara beradigan bir nechta usullaridan, xususan mashinali o‘qitishing klassik algoritmlari: *k* ta yaqin qo‘shnilar (KNN), tayanch vektorlar usuli (SVM), sodda bayes (Naive Bayes) va tasodifiy o‘rmon (Random Forest) algoritmlari ishlatilgan. Bundan tashqari ishda neyron tarmoq modellaridan rekurrent neyron tarmoqlarning turli xillari, hamda konvalyusion neyron tarmoq va rekurrent neyron tarmoqlarning gibrid modellaridan foydalanilgan [8].

Modelni o‘qitish bosqichida to‘plangan ma’lumotlar asosida tanlangan algoritm o‘qitiladi (moslashtiriladi). Sinov va baholash bosqichida tizimning aniqlik darajasini baholash uchun modelni alohida ma’lumotlar to‘plamida qanday aniqlikda ishlayotganligi sinab ko‘riladi. Agar natija kerakli natijani bermasa masalani yechish uchun boshqa giperparametrlar yoki boshqa algoritm tanlanadi.

Agar natija qo‘yilgan talabga javob beradigan bo‘lsa model yaratiladi. Tayyor bo‘lgan modelni amaliyotga tadbiq etish mumkin, ya’ni yangi matn kiritib natijani olish mumkin bo‘ladi.

Bu erda masalaning optimal yechimini topish uchun MO‘ algoritmlari va ular yordamida olingan natijalar muhim o‘rin tutadi. Shuning uchun matn qaysi tilda ekanligini aniqlash tajribalari quyidagi algoritmlar asosida amalga oshiriladi.



2-rasm. Matn tilini aniqlashda foydalanilgan algoritmlar

Yuqoridagi 2-rasmda keltirilgan algoritm-larning har birini ishlash tamoyillariga alohida qisqacha to‘xtalib o‘tiladi [6-8].

Sodda Bayes algoritmi. Sodda (Naïve) Bayes algoritmi NLP da qo‘llaniladigan oddiy, ammo samarali algoritm hisoblanadi [5]. Uning asosiy tamoyili shundaki, har bir xususiyat o‘zini boshqalaridan o‘zgarmaydigan (Naive) sifatda qabul qiladi. Bu algoritm Bayes teoremasidan foydalangan holda matnlarni sinflashtirishda va boshqa ba’zi vazifalarda yuqori natijalar olish uchun samarali hisoblanadi.

Bayes teoremasi, biror voqyea bo‘lganda noma’lum parametrlar (masalan, xususiyatlar)

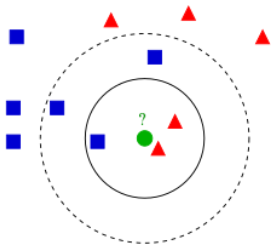
haqida ma'lumot berishni talab etadigan modellarda foydalaniladi.

Sodda Bayes algoritmi X ning shartlariga qarab Y sinfini aniqlashda quyidagi 1-formuladan foydalanadi [5]:

$$P\left(\frac{y}{x_1, x_2, \ldots x_n}\right) = p(y) * P\left(\frac{x_1}{y}\right) * P\left(\frac{x_2}{y}\right) * \ldots * P\left(\frac{x_n}{y}\right), \quad (1)$$

bunda, $X_1, X_2, \ldots X_n$ lar X ning har bir shartini ifodalaydi va bizga berilgan Y sinfi bo'yicha umumiy X ning ehtimolini topishga yordam beradi.

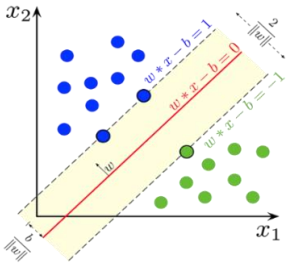
K ta yaqin qo'shnilar. K ta yaqin qo'shnilar (KNN) algoritmi yangi kuzatishlarni tasniflash yoki regressiya qilish uchun ishlatiladi. U ishlash jarayonida yangi namuna va o'quv ma'lumotlar to'plamidagi barcha namunalar o'rtasidagi masofani hisoblaydi, so'ngra eng yaqin K ta qo'shnisini tanlaydi va ularning sinflariga qarab ovoz berish yoki qiymatlarini o'rtachalashtirish orqali natijani aniqlaydi.



3-rasm. K ta yaqin qo'shnilar

KNN oddiy va samarali bo'lsada, katta ma'lumotlar to'plamlarida sekin ishlashi mumkin va K ning to'g'ri tanlanishini talab qiladi.

Tayanch vektorlar usuli (SVM). Tayanch vektorlar usuli (SVM- Support Vector Machine) tasniflash algoritmi ma'lumotlarni ikki sinfga ajratish uchun eng yaxshi ajratuvchi chiziq yoki gipertekislikni topishga harakat qiladi. Bu chiziq ikkala sinf o'rtasidagi eng katta masofani (margine) ta'minlaydi, shuningdek eng yaqin ma'lumot nuqtalari (support vectors) orqali margin masofasi maksimal qilinadi.

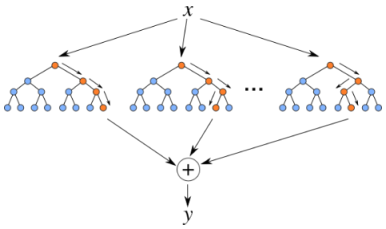


4-rasm. Tayanch vektorlar usuli

Agar ma'lumotlar chiziqli ajratilmasa, SVM "kernel trick" yordamida ma'lumotlarni yuqori o'lchamli fazoga o'tkazib, chiziqli ajratishni amalga oshiradi.

Bu usul kuchli umumlashish qobiliyatiga ega va turli sinflarga ajratish masalalarida samarali ishlaydi.

Tasodiy o'rmon. Tasodifiy o'rmon (Random Forest) algoritmi ko'p sonli qaror daraxtlarni o'z ichiga oladi va ularning natijalarini birlashtirib (ansambil usuli) tasniflash yoki regressiya qiladi. Har bir daraxt mustaqil ravishda o'quv to'plamining tasodifiy qismini va xususiyatlarning tasodifiy tanlangan kichik to'plamini ishlatib quriladi.

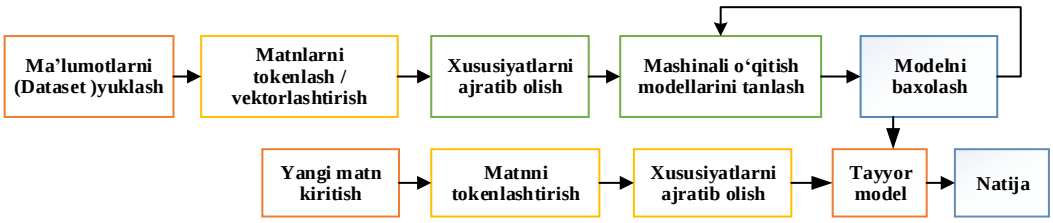


5-rasm. Tasodiy o'rmon

Natijada, tasniflashda har bir daraxtning ovoz berishi bilan eng ko'p ovoz olgan sinf tanlanadi.

Bu usul yuqori aniqlikni ta'minlaydi va haddan tashqari moslashuvchanlik (overfitting) xavfini kamaytiradi.

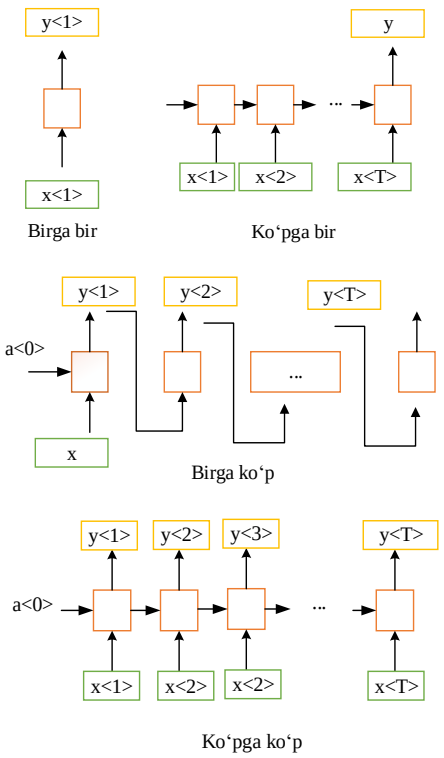
Ko'rib o'tilgan barcha sinflashtirish algoritmlaridan foydalanib matn tilini aniqlash ketma-ketligini quyida keltirilgan 6-rasmdagidek qurish mumkin.



6-rasm. Matn tilini aniqlash jarayoni blok-sxemasi

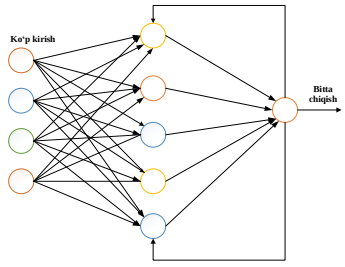
Takroriy neyron tarmoq (RNN). Takroriy neyron tarmoqlar (RNN) - ma'lumotlar ketma-ketligini qayta ishlash uchun mo'ljallangan sun'iy neyron tarmoq turi. Ular vaqtli ketma-ketlik ma'lumotlari, ovoz, tabiiy til va boshqa amallar kabi ketma-ketlikni talab qiladigan ishlar uchun yaxshi ishlaydi [6]. Uning turli variantlari mavjud bo'lib, RNN tarmoqlari bir biridan kiruvchi va chiqivchi qiymatlarning soni bo'yicha farqlanadi (7-rasm).

Maskur masala uchun RNN ning ko'pga bir arxitekturasiga mos keladi. Chunki, neyron tarmoqda kiruvchi ma'lumotlar matn (ko'p o'lchovli) va chiquvchi qiymati sinf indeksi (bitta qiymat) hisoblanadi. Ushbu RNN kirishlar ketma-ketligini oladi va bitta chiqish hosil qiladi.



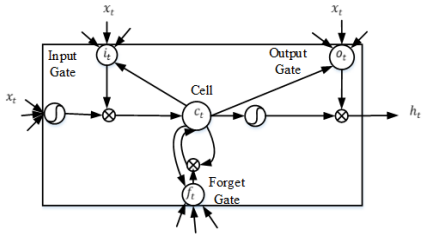
7-rasm. Takroriy neyron tarmolari turlari

Matnlarni sinflashtirish ushbu turdagi tarmoqning yaxshi namunasidir (8-rasm).



8-rasm. Takroriy neyron tarmoq arxitekturasi

Uzoq qisqa muddatli xotira (LSTM) neyron tarmoq modeli. Modellashtirish uchun qo'llaniladigan yana bir RNN ning bir turiga xotira bloklari deb nomlanadigan maxsus elementlardan tashkil topgan LSTM tarmog'i kiradi. Xotira bloklari tarmoq holatlarini vaqtincha saqlab turadigan yacheykalardan hamda geytlar deb nomlanadigan multiplikativ elementlar va axborot oqimlarini boshqarishdan tashkil topgan. Har bir xotira bloki kiruvchi va chiquvchi geytlardan, hamda yoddan chiqarish geytlaridan tashkil topgan. LSTM tarmog'ining xotira blokiga misol 9-rasmda ko'rsatilgan. Rasmdagi x_t t vaqt momentidagi kiruvchi vektor, h_t - chiquvchi vektor sifatida qarash mumkin [11].



9-rasm. LSTM tarmog'i xotira bloki

LSTM tarmoq yacheykasi uzoq vaqt davomida ma'lumotlarni saqlaydigan tarmoqning murakkab elementi sifatida qaraladi. Geytlar kirish ma'lumotlari qachon ahamiyatli ekanligini va uni yodda saqlashni yoki qachon ma'lumotlarni chiqishga yuborishni aniqlaydi (9-rasm).

LSTM tarmog'ini quyidagi ifodalar asosida ishlaydi.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i), \quad (2)$$

$$f_t = \sigma(W_{xt}x_t + W_{ht}h_{t-1} + W_{ct}c_{t-1} + b_t), \quad (3)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \quad (4)$$

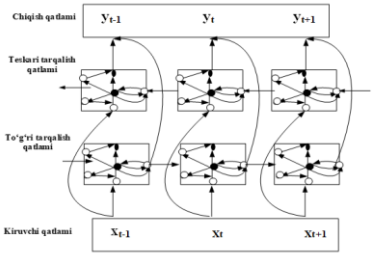
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o), \quad (5)$$

$$h_t = o_t \tanh(c_t), \quad (6)$$

bu erda σ sigmoit funksiya, i, f, o va c lar mos ravishda kiruvchi, yoddan chiqaruvchi, chiquvchi geytlar va holat yacheykalari sanaladi ularning hammasi yashirin vektor h bilan bir xil o'lchamga ega [10].

Ikki yo'nalishli uzoq qisqa muddatli xotira (BLSTM) neyron tarmoq modeli. RNN va LSTM lar afzalliklarini birlashtirgan holda ketma-ket modellashtirishni amalga oshirish uchun biz avvalgi uzoq muddatli kontekstli bog'liqliklardan, shuningdek keyingi vaqt qadamlaridan

foydadanishimiz mumkin. Bu jarayon 10–rasmdagi 3 ta *t*-1, *t*, *t*+1 vaqtli qadamlar uchun kengaytirilgan BLSTM modelida ko‘rsatilgan [13].



10-rasm. BLSTM ga asoslangan kengaytirilgan RNN model

Kirish to‘g‘ri va teskari LSTM qatlamlariga beriladi. Chiqish qatlami yoki keyingi yashirin qatlam (to‘g‘ri va teskari LSTM bosqichlari) LSTM ning to‘g‘ri va teskari bosqichlarini ulash orqali kirishni qabul qiladi. To‘g‘ri va teskari qatlamlar o‘rtasida yashirin aloqalar mavjud emas. Chiqish y_t uchun tenglama RNN bilan bir xil bo‘ladi.

Ushbu RNN tarmog‘i va uning modifikatsiyalangan turlari boshqa modellashtirishda qo‘llaniladigan algoritmlarga nisbatan quyidagi afzalliklarga va kamchiliklarga ega:

- Yutuqlari, RNN istalgan uzunlikdagi ketma-ketliklarga ishlov berish imkoniyatiga ega; RNN modeli har bir ma’lumotni vaqt davomida yodda saqlab qolish uchun modellashtiriladi, bu har qanday vaqtli qatorlarni bashoratlash uchun juda samarali hisoblanadi, bunda hatto kiruvchi o‘lcham katta bo‘lganda ham model o‘lchami oshmaydi; vaznlar vaqt qadamlarida tarqatilishi mumkin; RNN ixtiyoriy kiruvchi ma’lumotlarga ishlov berish uchun o‘zining ichki xotirasidan foydalanishi mumkin.

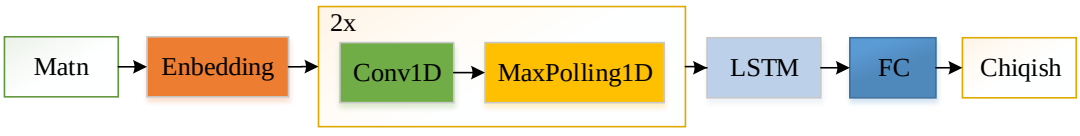
- Kamchiliklari, uning takrorlanuvchi tabiati hisobiga hisoblash sekin amalga oshiriladi; RNN modelni o‘qitish murakkab sanaladi; juda uzun ketma-ketliklarga ishlov berish murakkablashadi.

Umuman olganda RNN ni matnni tanish tizimining asosiy strukturasi bo‘lishiga imkon bermaydigan bir qancha kamchiliklarni keltirishimiz mumkin. Masalan, RNN larda matnning so‘zlar ketma-ketligini va vaqtli o‘zgaruvchanlikni aks ettiradigan mexanizmlari mavjud emas; RNN dinamikasi va topologiyasini belgilovchi bir qator parametrlar va bu parametrlarni hisoblash yoki tanlash uchun nazariy asos mavjud emas (ular ishlab chiquvchining xohishiga ko‘ra tanlanadi); O‘qitish jarayonini tezlashtiruvchi bir qator algoritmlar ishlab chiqilganligiga qaramay, u juda kuchli resurs talab qiladigan jarayon bo‘lib qolmoqda [16].

LSTM va konvolyusion neyron tarmog‘i (LSTM +CNN). Bu algoritm ikkita algoritm ketma ketligidan foydalangan holda ishlaydi. Bular LSTM va CNN algoritmlaridir. LSTM modellari, qisqa muddatli xotira yordamida uzun muddatli bog‘liq o‘zgaruvchilarni eslatib turadi va bunday ma’lumotni qayta ishlash jarayonlarida saqlash imkoniyatini beradi. Bu uzun va bog‘liq kontekstli matnlarni talab qiladigan vaziyatlarda foydalaniladi. LSTM modellari, bir matndagi so‘zlarning tashqi bog‘liqligini o‘rganib, matnning umumiy ma’nosini tushunishga yordam beradi.

CNN modellari, odatda ma’lumotlarini sinflashtirish uchun ishlatiladi, lekin matnlar uchun ham yaxshi natijalarni beradi [9].

Tadqiqot ishida yuqorida tar’iflab o‘tilgan mashinali o‘qishning klassik algortimlari va neyron tarmoq modellari asosida matn tilini aniqlash bo‘yicha qator tajriblar o‘tkazilib ko‘rildi. Buning natijasida klassik sinflashtirish algoritmlaridan sodda bayes va neyron tarmoq modellaridan gibrid CNN+LSTM modeli yuqori samaradorlik natijalarini qayd etdi. Shuning uchun ishda CNN+LSTM gibrid modeli loyihalash uchun tarmoq arxitekturasi va tarmoq giperparametrlari quyidagi 11-rasm va 1-jadvalda keltirib o‘tilgan.



11-rasm. CNN va LSTM neyron tarmoq modeli arxitekturasi

III. TAJRIBALAR

Tadqiqotni amalga oshirish ikki xil usulda shakllantirilgan o‘quv ma’lumotlar to‘plamida (dataset)dan foydalangan holda amalga oshirildi. Birinchi usulda belgilangan algoritmlardan foydalangan holda o‘zbek tili lotin yozuvida, o‘zbek tili kirill yozuvida, rus tili kirill yozuvida, rus tili lotin yozuvida, ingliz tili ingliz yozuvida, qoraqalpoq tili qoraqalpoq yozuvida, qoraqalpoq tili lotin yozuvuda berilgan matnning qaysi tilda yozilganligini aniqlash masalasi qarab chiqildi.

Ya'ni birinchi usulda ixtiyoriy yozuvga asoslangan holda amalga oshirildi.

Bu usulning mantig'iga ko'ra kiruvchi matn qaysi yozish turida bo'lishidan qat'iy nazar modeli uni aniqlashi zarur bo'ladi.

Datasetni shaklantirishda har bir sinf uchun 3500 tadan gaplar yig'ildi. Shakllantirilgan o'quv to'plamining tarkibi 2-jadvalda keltirib o'tilgan.

1-jadval. CNN va LSTM neyron tarmoq giperparametrlari

№	Qatlamning nomi	Qatlam parametrlari	Qatlamlar soni
1	Kiruvchi qatlam	So'zlar ketma-ketligi	1
2	Embedding qatlam	Qatlam o'lchami = [|V|,300] faollashtirish = ReLU,	1
3	Konvolyusion qatlam (Conv1D)	Filtr o'lchami = 1×5, soni = 128, faollashtirish = ReLU	2
4	Pulling qatlami	Pulling turi = MaxPolling1D, O'lchami = 1×5	
5	Rekurrent qatlam (LSTM)	Xotira yacheykalari soni = 128, faollashtirish = ReLU,	1
6	To'liq bog'lanishli qatlam (FC)	Qatlam o'lchami = 4, faollashtirish = softmax	1
7	O'qitish parametrlari	Optimallashtirish algoritmi: Adam, o'qitish qadami uzunligi = 0.001, Batch o'lchami = 30, Epochs = 100	

2-jadval. Birinchi usul yig'ilgan o'quv ma'lumotlarining taqsimlanishi

№	Ixtiyoriy yozuvga asoslangan matnlar	Jumlalar soni(ta)		
		O'quv	**Nazorat**	**Sinov**
1.	O'zbek tili lotin yozuvida	2800	350	350
2.	O'zbek tili kirill yozuvida	2800	350	350
3.	Rus tili kirill yozuvida	2800	350	350
4.	Rus tili lotin yozuvida	2800	350	350
5.	Ingliz tili	2800	350	350
6.	Qoraqalpoq tili	2800	350	350
7.	Qoraqalpoq tili lotin yozuvida	2800	350	350
Jami		**12222**	**250**	**350**

Ikkinchi usulda har bir tildagi yozilgan ma'lumotlarni lotin yozuvi shaklida ya'ni o'zbek tili, rus, ingliz va qoraqalpoq tillari uchun faqat lotin yozuviga asoslangan ma'lumotlardan foydalanilgan va har bir sinf ma'lumot turidan 3500 donadan gaplar shakllantirilgan.

Bu usulning mantig'iga ko'ra kiruvchi matn faqat lotin yozuvida bo'lishi zarur. Bunday holda model faqat lotin yozuvidagi ma'lumotlar bilan o'qitiladi. Sinov bosqichida maxus skript yatarilib kiruvchi matn qaysi yozuvda bo'lishidan qat'iy nazar uni lotin yozuvchiga o'g'irish amalga oshiriladi va modelga kiritiladi.

3-jadval. Ikkinchi usul yig'ilgan o'quv ma'lumotlarining taqsimlanishi

№	Lotin yozuvga asoslangan matnlar	Jumlalar soni(ta)		
		O'quv	**Nazorat**	**Sinov**
1.	O'zbek tili	2800	350	350
2.	Rus tili	2800	350	350
3.	Ingliz tili	2800	350	350
4.	Qoraqalpoq tili	2800	350	350
Jami		**11200**	**1400**	**1400**

Baholash ko'rsatkichlari. Matn tilini aniqlash sinflashtirish masalasi qatoriga kiradi. Shuning uchun baholash ko'rsatkichlarini shunga mos ravishda tanlanishi kerak. Buning uchun uchta ko'rsatkich ishlatilgan: aniqlik - Precision (7) da ko'rsatilgan, eslab qolish – Recall (8) da ko'rsatilgan va har bir yorliq uchun F1-ballari – F1-Score (9) da ko'rsatilgan. Ushbu ko'rsatkichlar quyidagicha aniqlanadi:

$$Precision\ (P) = \frac{TP}{TP + FP}, \quad (7)$$

$$Recall = \frac{TP}{TP + FN}, \quad (8)$$

$$F1\text{–}Score = \frac{2*TP}{2TP + FP + NP}. \quad (9)$$

Yuqoridagi (7) va (8) larda TP (True Positive) - haqiqiy ijobiy: model namunalarni ko‘rib chiqilayotgan sinfga to‘g‘ri tayinlashlar sonini. TN (True Negative) - haqiqiy salbiy: model namunalarni ko‘rib chiqilayotgan sinfga tegishli emasligini to‘g‘ri bildirishlari sonini. FP (False Positive) – noto‘g‘ri ijobiy: model namunalarni ko‘rib chiqilayotgan sinfga noto‘g‘ri tayinlashlar sonini va FN (False Negative) – noto‘g‘ri salbiy: model obyekt ko‘rib chiqilayotgan sinfga tegishli emasligini noto‘g‘ri bildirishlar sonini anglatadi [12].

Dastlabki tajribalar birinchi usulda yig‘ilgan matnning qaysi tilda ekanligini aniqlash uchun amalga oshirilgan. Ixtiyoriy yozuvga asoslangan holda amalga oshirilgan sinov natijalari 4-jadvalda keltirib o‘tilgan. Jadvalda precision, recall va f1-score aniqlik ko‘rsatkichlari sinov to‘plam namunalari uchun o‘rtacha qiymatlarini ifodalaydi.

4-jadval. Tanlangan algoritmlar asosida dastlabki sinov natijalari

Algoritmlar / Aniqlik	Sodda Bayes	LSTM	BLSTM	CNN +LSTM
Precision	96,42	94,84	95,79	97,16
Recall	96,87	95,71	96,51	97,47
F1-score	96,64	95,27	96,15	97,31
O‘rtacha	**96,64**	**95,27**	**96,15**	**97,31**

Olib borilgan dastlabki tajribalar shuni ko‘rsatdiki, ixtiyoriy yozuvga asoslangan holda matn tilini aniqlashni amalga oshirishda sinov natijasida yuqori natijalarni Sodda Bayes va CNN+LSTM gibrid modelari qayd etdi. Unga ko‘ra sinov to‘plamida Sodda Bayes algoritmi o‘rtacha aniqlik 96.64% ni va CNN+ LSTM gibrid modelida 97.31% tashkil etdi.

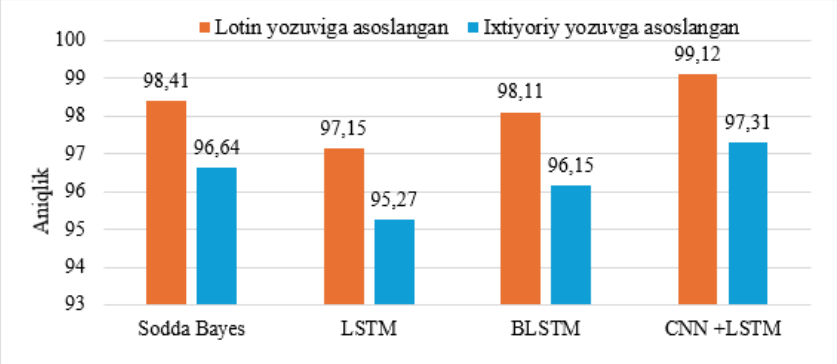
5-jadval. Tanlangan algoritmlar asosida ikkinchi sinov natijalari

Algoritmlar / Aniqlik	Sodda Bayes	LSTM	BLSTM	CNN +LSTM
Precision	98,83	97,37	98,35	99,32
Recall	97,99	96,93	97,87	98,93
F1-score	98,41	97,15	98,11	99,12
O‘rtacha	**98,41**	**97,15**	**98,11**	**99,12**

Ikkinchi tajribada berilgan ma’lumotlar lotin yozuviga o‘girgan holda amalga oshirildi. Ya’ni faqat lotin yozuvga asoslangan holda tariba o‘tkazildi. Sinov natijalari 5-jadvalda keltirib o‘tilgan.

Ikkinchi tajribalar shuni ko‘rsatdiki faqat lotin yozuvga asoslangan holda amalga oshirilgan sinov natijasida o‘rtacha aniqlik Sodda Bayes algoritmida 98.41%, LSTM algoritmida 97.15%, BLSTM algoritmida 98.11% ni va CNN+LSTM gibrid modeli yordamida sinovni amalga oshirish natijasi 99.12% aniqlikni ko‘rsatdi.

12-rasmda ko‘rib o‘tilgan ikkita yondashuvning sinov natijalari diagramma ko‘rishishida keltirilgan bo‘lib, qaysi yondashuv va qaysi algoritm matn tilini aniqlashda yaxshi natija berishini baholashi mumkin bo‘ladi. bunda, ikkinchi yondashuv ya’ni faqat lotin yozuviga asoslangan matnlarni sinflashtirish (matn tilini aniqlash) yuqoriroq natija ko‘rsatmoqda. Buning asosiy sababi sifatida sinflar sonining kamayishini keltirish mumkin. Birinchi yondashuvda sinflar soni 7 ta, ikkinchisida esa 4 ta. Bu esa modelning aniqroq bashorat qilish ehtimolini oshiradi. Yuqoridagi diagrammadan yana shuni ko‘rish mumkinki, sodda bayes mashinali o‘qitish algoritmi chuqur neyron tarmoqlariga asoslangan CNN-LSTM modellariga yaqin aniqlik natijasini ko‘rsatmoqda. Sodda bayes algoritmi soddaligi, hisoblash murakkabligining quyiligi hamda model hajmining kichikligi jihatidan neyron tarmoqlarga asoslangan modellardan ustunlik qiladi. Shuning uchun matn tilini aniqlashda sodda bayes algoritmidan ham foydalanish mumkin.



12-rasm. Ikkita usul bo‘yicha aniqliklarni solishtirish diagrammasi

IV. XULOSA

Tadqiqot ishida matnni qaysi tilda ekanligini aniqlash masalasi uchun ikki turdagi yondashuv taklif etilgan. Olib borgan tadqiqotlarimiz natijalari shuni ko‘rsatadiki, ixtiyoriy yozuv turiga asoslangan matnli ma’lumotdan foydalgan holda o‘qitilgan modelga nisbatan faqat lotin yozuviga asoslangan matnli ma’lumotlarda o‘qitilgan model yuqoriroq natijaga olib keladi.

Tajribalarda faqat lotin yozuvga asoslangan holda amalga oshirilgan sinov natijalari o‘rtacha aniqlik Sodda Bayes algoritmida 98.41%, LSTM algoritmida 97.15%, BLSTM algoritmida 98.11% ni va CNN+LSTM gibrid modeli yordamida sinovni amalga oshirish natijasi 99.12% aniqlikni ko‘rsatdi.

ADABIYOTLAR

[1] *Jurafskiy, D., & Martin, J. H.* "Speech and Language Processing" (3rd ed.) (2019).

[2] *Xolmatov O.A., Kamolov R.K.* NLP da savol-javob tizimlarini yaratish turlari va bosqichlari. Muhammad al-Xorazmiy nomidagi TATU Samarqand filiali. “O‘zbek tilining milliy korpusi: muammolar va vazifalar” mavzusidagi xalqaro ilmiy-amaliy konferensiya.

[3] *Xujayarov I.Sh, Ochilov M.M.* Neyron tarmoqlariga asoslangan nutq signallarini akustik modellashtirish usullari tahlili.

[4] *S. Ibragimova, T. Boburkhon, M. Abdullayeva*. Solving the problems of normalization of non-standard words in the text of the uzbek language Acta of Turin Polytechnic University in Tashkent 13 (3), pp. 38-42.

[5] *Bird, S., Klein, E., & Loper, E.* "Natural Language Processing with Python." O‘Reilly Media (2009).

[6] *Nielsen, M. A.* "Neural Networks and Deep Learning." Determination Press (2015).

[7] *Brownlee, J.* "Long Short-Term Memory Networks with Python." Machine Learning Mastery (2018).

[8] *Gulli, A., & Pal, S.* "Deep Learning with Keras." Packt Publishing (2017).

[9] *Zhang, J., Zhao, J., & LeCun, Y.* "Character-level Convolutional Networks for Text Classification." Advances in Neural Information Processing Systems (2015).

[10] *Hochreiter S., Schmidhuber J.* Long short-term memory. Neural computation. 1997. vol. 9. no. 8. pp. 1735–1780.

[11] *Musaev M., Khujayorov I., Ochilov M.* “Development of integral model of speech recognition system for Uzbek language” IEEE 14th International Coference on Application of Information and Communication Technologies (AICT). 07-09 October 2020.

[12] *Musaev M., Khujayarov I., Ochilov M.* “Speech Recognition Technologies Based on Artificial Intelligence Algorithms” Intelligent Human Computer Interaction: 14th International Conference, IHCI 2022, Tashkent, Uzbekistan, October 20–22, 2022, Pages 51–62.

[13] *Abdullaeva M.I., Juraev D.B., Ochilov M.M., Rakhimov M.F.,* Uzbek Speech Synthesis Using Deep Learning Algorithms. The 14th International Conference on Intelligent Human Computer Interaction, Springer, (LNCS,volume 13741), Tashkent – 2023, pp 39–50.

[14] *Xujayarov I.Sh, Ochilov M.M.* Neyron tarmoqlariga asoslangan nutq signallarini akustik modellashtirish usullari tahlili.

[15] *S Ibragimova, T Boburkhon, M Abdullayeva.* Solving the problems of

normalization of non-standard words in the text of the uzbek language Acta of Turin Polytechnic University in Tashkent 13 (3), pp. 38-42 .
[16] *Abdullayeva M.I., Jurayev D.B, Ochilov M.M.* Nutqni imo-ishora tiliga tarjima qilish tizimlarida matnlarga ishlov berish. TATU xabarlari, b. 112-118.

Поступила в редакцию 06.04.2024

Citation: *Xujayarov, I., Ochilov, M., Xolmatov, O., & Jurayev, D. (2024).* Sun'iy intellekt algoritmlari asosida matn tilini avtomatik aniqlash. Международный Журнал Теоретических и Прикладных Вопросов Цифровых Технологий, 7(2), 59–67. https://doi.org/10.62132/ijdt.v7i2.182

AUTOMATIC RECOGNITION OF TEXT LANGUAGE BASED ON ARTIFICIAL INTELLIGENCE ALGORITHMS

Khujayarov I.Sh.[1], Ochilov M.M.[2], Kholmatov O.A.[2], Juraev D.B.[2]
[1] Samarkand branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Samarkand, Uzbekistan
[2] Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan
i.khujayorov@tuit.uz, ochilov.mannon@mail.ru, xolmatov.orzumurod@gmail.com, dilsamtuit@tuit.uz

Abstract. *The article discusses approaches to text recognition based on artificial intelligence algorithms. Since text language identification is a classification problem in artificial intelligence, the article examines the capabilities of several classification algorithms using machine learning and neural network models. The study addresses the issue of forming a training dataset for a model that identifies Uzbek, English, Russian, and Karakalpak languages. Additionally, a comparative analysis of the accuracy indicators of the models used for text language identification is conducted.*

Keywords: *NLP, crowling, text language, training dataset, machine learning, classification, deep learning, recurrent neural networks, model accuracy evaluation.*

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ ЯЗЫКОВОГО ТЕКСТА НА ОСНОВЕ АЛГОРИТМОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Хужаяров И.Ш.[1], Очилов М.М.[2], Холматов О.А.[2], Жураев Д.Б.[2]
[1] Самаркандский филиал Ташкентского университета информационных технологий имени Мухаммада ал-Хорезми, Самарканд, Узбекистан
[2] Ташкентский университет информационных технологий имени Мухаммада ал-Хорезми, Ташкент, Узбекистан
i.khujayorov@tuit.uz, ochilov.mannon@mail.ru, xolmatov.orzumurod@gmail.com, dilsamtuit@tuit.uz

Аннотация. *В статье рассматриваются подходы к распознаванию текста на основе алгоритмов искусственного интеллекта. Поскольку идентификация языка текста является проблемой классификации в искусственном интеллекте, в статье рассматриваются возможности нескольких алгоритмов классификации с использованием моделей машинного обучения и нейронных сетей. Рассмотрен вопрос формирования обучающего набора данных для модели, определяющей узбекский, английский, русский и каракалпакские языки. Дополнительно проводится сравнительный анализ показателей точности моделей, используемых для идентификации языка текста.*

Ключевые слова: *Обработка естественного языка-NLP, краулинг, текстовый язык, обучающий набор данных, машинное обучение, классификация, глубокое обучение, рекуррентные нейронные сети, оценка точности модели.*