

O'ZBEK TILI UCHUN TABIIY TILNI QAYTA ISHLASHDA (NLP) MASHINALI O'RGANISHNING (MACHINE LEARNING) O'RNI

Sharipov Maksud Siddiqovich,

Urganch davlat universiteti dotsenti

maqsbek72@gmail.com

Adinayev Xushnudbek Saylboyevich,

TATU Urganch filiali assistant o'qituvchi

hushnud8414@mail.ru

Raximov Raximjon Rajapboyevich,

Urganch davlat universiteti katta o'qituvchisi

rahimjon03@mail.ru

Annotatsiya. Globalizatsiya va raqamli axborotlashtirish jarayonida o'zbek tilining rivojida axborot texnologiyalari o'ta muhim o'ringa ega. Kompyuter lingvistikasining tabiiy tilni qayta ishlash (Natural Language Processing – NLP) jarayoni va tabiiy tilning kompyuter tushunadigan til (formal) shakli yaratilishi natijasida tilga oid masalalar (tahrir, tahlil, tarjima, elektron matnni ovozlashtirish, og'zaki nutqni elektron matnga aylantirish, robot bilan muloqot qilish, yirik matnni ixcham matnga aylantirish kabi) bo'yicha kompyuter dasturlari yaratilmoqda.

Machine Learning va tabiiy tilni qayta ishlash (Natural Language Processing– NLP) so'nggi paytlarda mashhurlikka erishgan sun'iy intellektning muhim sohalaridir. Machine Learning va tabiiy tilni qayta ishlash sun'iy agentni, sun'iy aqlli agentga aylantirishda juda muhim rol o'ynaydi. Sun'iy ravishda aqlli tizim atrof-muhitdan yaxshiroq ma'lumotlarni qabul qilishi va tabiiy tilni qayta ishlashning rivojlanishi tufayli atrof-muhitga foydalanuvchilar uchun qulay tarzda ta'sir qilishi mumkin.

Kalit so'zlar: *tabiiy tilni qayta ishlash, NLP, Machine Learning, Deep Learning, NLTK.*

Abstract. In the process of globalization and digital informatization, information technologies play a very important role in the development of the Uzbek language. As a result of the process of natural language processing (NLP) of computer linguistics and the creation of a language (formal) form that can be understood by a computer, language-related issues (editing, analysis, translation, voicing of electronic text, converting spoken speech into electronic text, using robot computer programs are being created for communicating, converting large text into compact text).

Machine Learning and Natural Language Processing (NLP) are important areas of artificial intelligence that have gained popularity recently. Machine learning and natural language processing play a very important role in transforming an artificial agent into an artificially intelligent agent. An artificially intelligent system can receive better information from the environment and interact with the

environment in a user-friendly way thanks to the development of natural language processing.

Keywords: *natural language processing, NLP, Machine Learning, Deep Learning, NLTK.*

Аннотация: В процессе глобализации и цифровой информатизации информационные технологии играют очень важную роль в развитии узбекского языка. В результате процесса обработки естественного языка (НЛП) компьютерной лингвистики и создания языковой (формальной) формы, понятной компьютеру, даны ответы на вопросы, связанные с языком (редактирование, анализ, перевод, озвучивание электронного текста, преобразование устной речи в электронный текст, с помощью роботов создаются компьютерные программы для общения, преобразования большого текста в компактный текст).

Машинное обучение и обработка естественного языка (NLP) — важные области искусственного интеллекта, которые в последнее время приобрели популярность. Машинное обучение и обработка естественного языка играют очень важную роль в превращении искусственного агента в агента с искусственным интеллектом. Система с искусственным интеллектом может получать более качественную информацию из окружающей среды и взаимодействовать с окружающей средой удобным для пользователя способом благодаря развитию обработки естественного языка.

Ключевые слова: *обработка естественного языка, NLP, Machine Learning, Deep Learning, NLTK.*

Jahon tilshunosligida XX asrning 50-yillaridan til va matnni avtomatik tahlil qilish muammolari bilan bog'liq masalalar kompyuter texnologiyalari yordamida hal qilinmoqda. Natijada matnni avtomatik tushunish, mashina tarjimasini, matnni referatlash (gipermatnni asosiy mazmunni o'zida saqlagan minimatnga keltirish, ya'ni matn hajmini qisqartirish, uni qisqa bayon holiga aylantirish), tasniflash mavzu, uslub va janr jihatidan o'zaro yaqin matnlarni guruhlash), matnni tahrir va tahlil qilish, matnni generatsiyalash (bir nechta tabiiy tilga oid hujjatli matnlardan formula, texnik ishlanmalar, dasturiy tizimlarni yig'ish), matndan ma'lumotni olish, og'zaki nutqni raqamli ma'lumotda berish va aksincha, yozma nutqni og'zaki nutqqa aylantirish, tabiiy tilning Milliy korpusi va boshqa tur lingvistik korpuslarini yaratish, ontologik lug'atlar bazasini shakllantirish kabi dolzarb masalalar kompyuter lingvistikasining mundarijasini egallagan va tabiiy tilni qayta ishlash (Natural Language Processing / NLP) sohasida asosiy vazifa hamda yo'nalishlariga aylangan

Tabiiy tilni qayta ishlash (NLP) nima?

Kompyuterlar kundalik turmushning ajralmas qismi bo'lib, jadvalli / elektron ma'lumotlar bilan ishlashda juda qulay texnik vositasi hisoblanadi. Biroq, odamlar odatda jadvallar shaklida emas, balki so'zlar va jumlar bilan muloqot qilishadi. Og'zaki va yozma nutq matnlari xususiy bo'lib, til umumiylik kasb etib, muayyan

tuzilish (struktura)ga ega [1]. Shu bois kompyuterlar uchun ushbu turdagi ma'lumotlarni qayta ishlash usullarini yaratish taqozo etiladi. Tabiiy tilni qayta ishlash (NLP)dan maqsad kompyuterlarga strukturlangan matnni tushuntirish va mazmun olishni o'rgatishdan iborat. Tabiiy tilni qayta ishlash (NLP) sun'iy intellektning kichik sohasi bo'lib, uning maqsadi kompyuterlar va odamlar o'rtasidagi o'zaro aloqalarni o'rnatishdan iborat. Tabiiy tilni qayta ishlashga oid masalalarni yechish uchun Python dasturlash tilida yaratilgan NLTK kutubxonasi yuklab olinishi talab etiladi [2].

NLP yo'nalishlari

- Mashina tarjimasi (Machine Translation)
- Nutqni tanish (Speech Recognition)
- Kayfiyat / tuyg'ularni tahlil qilish (Sentiment Analysis)
- Savollarga javob berish (Question Answering)
- Matnni referatlash (Summarization of Text)
- Chatbot (Chatbot)
- Intellektual tizimlar (Intelligent Systems)
- Matn tasniflari (Text Classifications)
- Belgilarni aniqlash (Character Recognition)
- Imlo tekshiruvi (Spell Checking)
- Spamni aniqlash (Spam Detection)
- (Matnni) Avtomatik to'ldirish (Autocomplete)
- Nomga ega obyektning aniqlash (Named Entity Recognition)
- Taklifli terish (Predictive Typing)

Sun'iy intellekt tizimi qabul qilingan ma'lumotlarni qayta ishlashi va Machine Learning usullarini qo'llash tufayli uning harakatlari uchun yaxshiroq bashoratlarni amalga oshirishi mumkin.

Machine Learning tizimga o'tgan tajriba va misollardan o'rganish qobiliyatini beradi. Umumiy algoritmlar dasturlashtirilgan narsaga ko'ra qat'iy belgilangan ijro to'plamini bajaradi va ular noma'lum muammolarni hal qilish imkoniyatiga ega emas. Haqiqiy dunyoda duch kelinadigan muammolarning aksariyati an'anaviy algoritmlarning samaradorligini pasaytiradigan ko'plab noma'lum o'zgaruvchilarni o'z ichiga oladi. Bu yerda Machine Learning birinchi o'ringa chiqadi. Yuqorida sanab o'tilgan yo'nalishlarning barchasi machine learning yordamida modellarni qurish uchun o'z ma'lumotlar bazasi – datasetlarni talab qiladi. O'zbek tili uchun ham bunday datasetlardan ancha qurilgan, masalan Sentiment tahlil datasetlari [3], [4], so'zlar o'xshashlik va bog'liqlik darajalarini o'z ichiga olgan semantik tahlil dataseti [5], hamda o'zbek tilidagi nomuhim so'zlar datasetlari [6] shular qatoriga kiradi.

Keltirilgan klassik misollardan ba'zilar spam pochta ni aniqlashni o'z ichiga oladi. Elektron pochta ni qonuniy yoki spam ekanligini aniqlash va tasniflash ko'plab noma'lumlarni o'z ichiga oladi. Spam filtrlarini chetlab o'tishning ko'plab usullari mavjud. An'anaviy algoritm ishlashi uchun har bir xususiyat va o'zgaruvchi

qattiq kodlangan bo'lishi kerak, bu juda qiyin, agar iloji bo'lsa. Holbuki, mashinali o'qitish algoritmi umumiy qoidani o'rganish va shakllantirish qobiliyati tufayli bunday muhitda ishlashi mumkin. O'zbek tili uchun ham machine learning yordamida ancha ishlar qilingan, ulardan gap bo'laklari(parts-of-speech)ni ajratib beruvchi dataset va dasturiy vositasi [7], O'zbek tilidagi nomuhim so'zlarni topish metodi [8], hamda gaplar o'zagini aniqlashda ishlatiladigan stemming dasturiy vositalari [9] shular jumlasidandir.

Deep Learning - bu Machine Learning algoritmlarining ixtisoslashuvi, sun'iy neyron tarmog'idir. So'nggi paytlarda Deep Learning usullari keng qo'llanilib, yaxshi natijalarga erishayotgani kuzatilmoqda.

Arxitektura bo'yicha qaror qabul qilishda Deep Learning texnikasi tomonidan taqdim etilgan moslashuvchanlik ushbu texnikalar muvaffaqiyatining muhim sabablaridan biridir. Deep Learning usullari tabiiy tilni qayta ishlash bo'yicha tadqiqotlar uchun ishlatiladigan Machine Learning usullari orasida birinchi o'rinda turadi.

Tabiiy tilni qayta ishlash foydalanuvchilarning murakkab elektronika bilan o'zaro aloqasini osonlashtirgan bo'lsa ham, boshqa tomondan, bu aloqani amalga oshirishga imkon beradigan sahna ortida ko'plab qayta ishlash jarayoni sodir bo'ladi. Mashinali o'qitish tilni qayta ishlashda juda muhim rol o'ynadi.

Machine Learning tabiiy tilni to'g'ri qayta ishlashda rol o'ynashdan tashqari, tabiiy tilni qayta ishlashning muhim ilovalarida ham juda konstruktiv rol o'ynaydi. Hissiy tahlil, Chatbot tizimlari, Savollarga javob berish tizimlari, Axborot qidirish tizimlari, Mashina tarjimai va Elektron pochta tasniflash kabi muhim tabiiy tilni qayta ishlash ilovalari yaxshi ishlashi uchun Machine Learning usullarini o'z ichiga oladi.

Tabiiy tilni qayta ishlashda mashinani o'rganishning o'rni

Mashina tabiiy tilni tushunishi uchun tabiiy tilni qayta ishlash ko'p bosqichlarni qamrab oladi. Ushbu bosqichlar morfologik tahlil, sintaktik tahlil, semantik tahlil, nutq tahlili va pragmatik tahlilni o'z ichiga oladi, odatda, bu tahlil vazifalari ketma-ket qo'llaniladi. Machine Learning deyarli barcha jarayonlarda u yoki bu shaklda muhim qo'shimcha qiymat sifatida ishlaydi. Hozirgi kunda o'zbek tilida NLP masalalarini yechish bo'yicha ko'plab ilmiy ishlar olib borilmoqda.

1. Morfologik tahlil:

Yuqorida aytib o'tilganidek, hisoblash tizimi tomonidan olingan ma'lumotlar 0 va 1 lar shaklida bo'ladi. Bu 0 va 1 larni ASCII kodi yordamida alifboga aylantirish mumkin. Shunday qilib, aytish mumkinki, mashina unga jumla yoki paragraf berilganda bir nechta belgilarni oladi. Morfologik tahlil darajasida birinchi vazifa so'z va gaplarni aniqlashdir. Ushbu identifikatsiya tokenizatsiya deb ataladi. Tokenizatsiya uchun ko'plab turli xil Machine learning va Deep Learning algoritmlari, shu jumladan Yordam vektor mashinasi (Support Vector Machine) va takroriy neyron tarmog'i (Recurrent Neural Network) qo'llanilgan.

Tokenizatsiya tugallangandan so'ng mashinada bir nechta so'zlar va jumlar mavjud bo'ladi. Tuzilgan gaplarning aksariyati affikslardan iborat. Ushbu affikslar mashinalar ishini murakkablashtiradi, chunki barcha so'zlarni barcha mumkin bo'lgan affikslari bilan o'z ichiga olgan so'z ma'nosi lug'atiga ega bo'lish deyarli mumkin emas. Demak, morfologik tahlil darajasining navbatdagi vazifasi bu affikslarni olib tashlashdir. Bu affikslarni o'zaklash yoki lemmatizatsiya yordamida olib tashlash mumkin. Tasodifiy o'rmon (random forest) va qaror daraxti (decision tree) kabi Machine Learning algoritmlari stemming vazifasini bajarishda juda muvaffaqiyatli bo'ldi.

2. Sintaktik tahlil

Tabiiy tilni qayta ishlashning navbatdagi vazifasi berilgan gap tilning grammatik qoidasiga amal qilishini tekshirishdan iborat. Buning uchun so'zlar birinchi navbatda so'z turkumlariga ajratiladi. Bu sintaktik tahlilchilarga grammatika qoidalarini tekshirishda yordam beradi. Tasodifiy o'rmon va takroriy neyron tarmog'i kabi mashinali o'qitish va chuqur o'qitish algoritmlari ushbu vazifa uchun muvaffaqiyatli qo'llanildi. Sintaktik parserlarni amalga oshirish uchun K- eng yaqin qo'shni (K- nearest neighbor) kabi mashinani o'rganish algoritmlari ham ishlatilgan. Bu sohada o'zbek tili uchun ham qilingan eng katta ishlardan biri bu sintaktik tahlil qilingan datasetning yaratilganidir [10].

3. Semantik tahlil

Bu bosqichda so'z ma'nolari izohli lug'atlar yordamida aniqlanadi. Bu yerda duch kelinadigan muammo shundaki, bir xil so'z jumla kontekstiga ko'ra turli xil ma'nolarga ega bo'lishi mumkin. Masalan, “o't” so'zi olov yoki “o't” o'simlik ma'nosini anglatishi mumkin, bu noaniqlikni keltirib chiqaradi. Shunday qilib, ushbu noaniqlikni bartaraf qilish so'z ma'nosini ajratish deb ataladigan tabiiy tilni qayta ishlashning ushbu bosqichidagi muhim vazifalardan biridir.

So'z ma'nosini aniqlash - bu turli darajadagi muvaffaqiyatlar bilan tadqiq qilingan klassik tasniflash muammolaridan biridir. Tasodifiy o'rmon, gradientni kuchaytirish va qarorlar daraxtlari kabi mashinali o'qitish usullari muvaffaqiyatli qo'llanildi. Biroq, so'nggi paytlarda takroriy neyron tarmog'i, uzoq qisqa muddatli xotiraga asoslangan takroriy neyron tarmoq, eshikli takrorlanuvchi birlikka asoslangan takroriy neyron tarmoq va konvolyutsiya neyron tarmog'i kabi chuqur o'qitish algoritmlari o'rganilgan va juda yaxshi natijalar bergan.

4. Diskurs tahlili

Tahlilning joriy ko'rinishidan tashqarida olmoshlar qo'llanilgan yoki ma'lum mavzular/obyektlarga havola qilingan holatlar mavjud. Bunday hollarda semantik tahlil gapga kerakli ma'no bera olmaydi. Bu mashinali o'qitish va chuqur o'qitish algoritmlari tomonidan hal qilingan mos yozuvlar yechimining yana bir klassik muammosidir.

5. Pragmatik tahlil

Ko'p vaqtlarda jumlar so'zlar tasvirlagandan ko'ra chuqurroq ma'noni anglatadi. Ya'ni, mashina semantik tahlildan so'ng tushunilgan so'z ma'nosidan voz

kechib, mo'ljallangan yoki nazarda tutilgan ma'noni egallashi kerak. Buni qilishdan ko'ra aytish osonroq. Ko'p yillar davomida bu tabiiy til jarayoni tadqiqotchilarni qiziqtirib kelmoqda. Pragmatik tahlilning klassik misollaridan biri kinoyani aniqlashdir.

Ko'pchilik, aslida deyarli barcha turli xil Machine learning va Deep learning algoritmlari kinoyani aniqlash yoki umuman pragmatik tahlilni amalga oshirish uchun turli xil muvaffaqiyatlar bilan qo'llanilgan.

Xulosa. Shunday qilib, Machine Learning va Deep Learning usullari tabiiy tilni qayta ishlash sohasida foydalanish uchun keng qamrovli tadqiq qilinayotganini kuzatish mumkin. Ko'rinib turibdiki, bu o'rganish usullari deyarli barcha tabiiy til vazifalarini qayta ishlashda, shuningdek, tabiiy tilni qayta ishlashning deyarli barcha ilovalarida muhim rol o'ynaydi.

Tabiiy til vazifalarini har xil qayta ishlash va tabiiy tilni qayta ishlashning turli xil ilovalari o'z-o'zidan tadqiqotning turli sohalaridir. Va hozirda ushbu tadqiqot sohalarida Machine Learning va Deep Learning usullari yuqori darajadagi muvaffaqiyat bilan keng ko'lamda o'rganilmoqda. Xulosa qilib aytish mumkinki, Machine Learning va Deep Learning texnikasi tabiiy tilni qayta ishlash va uning ilovalarida juda ijobiy rol o'ynamoqda.

Foydalanilgan adabiyotlar:

1. V. Garousi, S. Bauer, and M. Felderer, “NLP-assisted software testing: A systematic mapping of the literature,” *Information and Software Technology*, vol. 126. 2020. doi: 10.1016/j.infsof.2020.106321.
2. DataCamp, “spaCy,” Python Cheat Sheet, 2020.
3. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodriguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers, 2022*, pp. 232–243.
4. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings, 2022*, vol. 3315, pp. 126–136. [Online]. Available: www.scopus.com
5. U. Salaev, E. Kuriyozov, and C. Gómez-Rodriguez, “SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language,” in *1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings, 2022*, pp. 199–206. [Online]. Available: www.scopus.com
6. K. Madatov, S. Bekchanov, and J. Vičič, “Lists of uzbek stopwords,” *Univerza na Primorskem, Inštitut Andrej Marušič*, 2021.



7. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, "UzbekTagger: The rule-based POS tagger for Uzbek language," arXiv preprint arXiv:2301.12711, 2023.
8. X. Madatov, M. Sharipov, and S. Bekchanov, "O'ZBEK TILI MATNLARIDAGI NOMUHIM SO'ZLAR," COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS, vol. 1, no. 1, 2021.
9. M. Sharipov and U. Salaev, "Uzbek affix finite state machine for stemming," arXiv preprint arXiv:2205.10078, 2022.
10. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, "Creating a morphological and syntactic tagged corpus for the Uzbek language," arXiv preprint arXiv:2210.15234, 2022.