

V SHO‘BA. KOMPYUTER LINGISTIKASIDA SUN‘IY INTELLEKT

O‘ZBEK TILIDA SIMPTOMLAR ASOSIDA KASALLIKNI ANIQLASH MODELI

Elov Botir Boltayevich

Texnika fanlari doktori (DSc), dotsent

elov@navoiy-uni.uz

ToshDO‘TAU

Ro‘zimurodova Jasmina Toshpo‘lot qizi

Kompyuter lingvistikasi yo‘nalishi talabasi

ruzimuradovajasmina@gmail.com

ToshDO‘TAU

Annotatsiya. Mazkur maqolada o‘zbek tilidagi bemor shikoyatlari va simptomlari asosida kasalliklarni aniqlashga mo‘ljallangan kompyuter modeli ishlab chiqish masalasi ko‘rib chiqiladi. Tadqiqotning asosiy maqsadi o‘zbek tilidagi tibbiy matnlarni avtomatik tahlil qilish va bemor simptomlari asosida ehtimoliy kasalliklarni aniqlash usullarini ishlab chiqishdan iborat. Tadqiqot jarayonida tabiiy tilni qayta ishlash (Natural Language Processing, NLP) texnologiyalari, matnlarni oldindan qayta ishlash usullari hamda mashinaviy o‘rganish algoritmlaridan foydalaniladi. Matnlardan simptomlarni aniqlash, ularni tasniflash va kasalliklarni klassifikatsiya qilish uchun maxsus model taklif etildi. Taklif etilayotgan model bemor shikoyatlaridan muhim tibbiy ma’lumotlarni ajratib olish hamda simptomlar asosida kasalliklarni aniqlash orqali samarali natijalar beradi. Mazkur tadqiqot o‘zbek tilida tibbiy matnlarni avtomatik tahlil qilish va intellektual diagnostika tizimlarini yaratish uchun ilmiy asos bo‘lib xizmat qiladi.

Kalit so‘zlar: *kompyuter lingvistikasi, tabiiy tilni qayta ishlash, tibbiy matnlar, simptomlar, kasallikni aniqlash, sun‘iy intellekt, mashinali o‘qitish.*

Abstract. This article examines the development of a computer model designed to identify diseases based on patient complaints and symptoms in the Uzbek language. The main objective of the study is to develop methods for

automatically analyzing Uzbek-language medical texts and identifying potential diseases based on patient symptoms. During the research process, Natural Language Processing (NLP) technologies, text preprocessing methods, and machine learning algorithms are used. A specialized model has been proposed for extracting symptoms from texts, classifying them, and diagnosing diseases. The proposed model delivers effective results by extracting important medical information from patient complaints and diagnosing diseases based on symptoms. This research serves as a scientific basis for creating automatic analysis of medical texts and intelligent diagnostic systems in the Uzbek language.

Keywords: computer linguistics, natural language processing, medical texts, symptoms, disease diagnosis, artificial intelligence, machine learning.

Kirish. So‘nggi yillarda sun‘iy intellekt va tabiiy tilni qayta ishlash texnologiyalarining rivojlanishi tibbiyot sohasida yangi imkoniyatlarni yaratdi. Ayniqsa, bemorlarning shikoyatlari va simptomlari asosida kasalliklarni aniqlash diagnostika jarayonini tezlashtirish va shifokorlarga yordam berish nuqtai nazaridan muhim ahamiyat kasb etadi. Zamonaviy tibbiy axborot tizimlarida bemorlar ko‘pincha o‘z shikoyatlarini matn ko‘rinishida ifodalaydilar. Ushbu matnlarni avtomatik ravishda tahlil qilish va ulardan muhim klinik ma‘lumotlarni ajratib olish uchun tabiiy tilni qayta ishlash texnologiyalaridan foydalaniladi. Tibbiy matnlarni avtomatik tahlil qilish, ulardan simptomlarni aniqlash va kasalliklarni tasniflash masalasi so‘nggi yillarda ko‘plab ilmiy tadqiqotlarning markazida bo‘lib kelmoqda. Xususan, klinik matnlarni qayta ishlash uchun turli xil algoritmlar va chuqur o‘rganish modellaridan foydalanish samarali natijalar berayotgani qayd etilgan. Bunday tizimlar shifokorlarga diagnostika jarayonida yordam berish, tibbiy ma‘lumotlarni tezkor tahlil qilish va bemorlar uchun maslahat tizimlarini yaratishda qo‘llanilishi mumkin. Dunyo miqyosida tibbiy matnlarni tahlil qilish uchun bir qator sun‘iy intellekt modellaridan foydalanilmoqda. Jumladan, BERT, BioBERT va

ClinicalBERT kabi transformer modellar klinik matnlarni semantik tahlil qilishda yuqori natijalar ko'rsatgan. Ushbu modellar yordamida bemor shikoyatlaridan simptomlarni aniqlash, kasalliklarni tasniflash va diagnostika jarayonini avtomatlashtirish mumkin. Biroq ushbu tadqiqotlarning aksariyati ingliz tilidagi klinik matnlarga qaratilgan bo'lib, kam resursli tillar, xususan o'zbek tilida tibbiy matnlarni avtomatik tahlil qilish bo'yicha tadqiqotlar yetarli darajada olib borilmagan. O'zbek tilidagi tibbiy matnlar o'ziga xos lingvistik xususiyatlarga ega. Tilning agglutinativ tuzilishi sababli so'zlar turli grammatik qo'shimchalar yordamida o'zgaradi. Jumladan, bir xil simptom turli shakllarda ifodalanishi mumkin:

“bosh og'rig'i”, “boshim og'riyapti”, “boshim qattiq og'riyapti”

kabi ifodalar bir xil klinik simptomni bildiradi. Bu holat matnlarni avtomatik tahlil qilish jarayonini murakkablashtiradi va maxsus lingvistik hamda hisoblash usullarini qo'llashni talab etadi.

Shu bilan birga, o'zbek tilidagi tibbiy matnlar ko'pincha erkin nutq shaklida yoziladi. Bemorlar o'z shikoyatlarini ifodalashda tibbiy terminologiyadan emas, balki kundalik nutq birliklaridan foydalanadilar. Masalan, *“qornim achishyapti”, “ichim kuyayapti”, “ovqat hazm bo'lmayapti”* ga o'xshash ifodalar tibbiy nuqtai nazardan turli kasalliklar bilan bog'liq simptomlarni bildirishi mumkin. Bunday ifodalarni avtomatik ravishda aniqlash va ularni standart tibbiy terminologiyaga moslashtirish murakkab lingvistik va hisoblash muammolarini keltirib chiqaradi. Tibbiy matnlarni qayta ishlash jarayonida yana bir muhim muammo sinonim birliklar va turli variantlarda ifodalangan simptomlarni aniqlash bilan bog'liq. Misol tariqasida, *“isitma”, “harorat ko'tarilishi”, “tana harorati oshishi”* kabi birliklar semantik jihatdan bir xil simptomni ifodalashi mumkin. Shuning uchun bunday birliklarni aniqlash va ularni yagona klinik tushunchaga birlashtirish tabiiy tilni qayta ishlash tizimlari uchun muhim vazifa hisoblanadi.

Bundan tashqari, tibbiy matnlarni tahlil qilish jarayonida kontekstni to'g'ri aniqlash ham muhim ahamiyatga ega. Bemor shikoyatlarida bir nechta simptomlar bir vaqtning o'zida uchrashi mumkin. *“Boshim og'riyapti va ko'nglim ayniyapti”* kabi ifodalar bir nechta klinik simptomlarni o'z ichiga oladi. Bunday holatlarda simptomlar o'rtasidagi bog'liqlikni aniqlash va ular asosida ehtimoliy kasalliklarni aniqlash uchun murakkab tahlil usullaridan foydalanish zarur. Zamonaviy tibbiy axborot tizimlarida katta hajmdagi matnli ma'lumotlar to'planmoqda. Elektron tibbiy yozuvlar, bemor shikoyatlari, diagnostika hisobotlari va tibbiy maslahatlar ko'pincha matn ko'rinishida saqlanadi. Ushbu ma'lumotlarni qo'lda tahlil qilish ko'p vaqt talab qiladi va inson omiliga bog'liq xatolar yuzaga kelishi mumkin. Shu sababli tibbiy matnlarni avtomatik tahlil qilishga mo'ljallangan intellektual tizimlarni yaratish muhim ilmiy va amaliy vazifalardan biri hisoblanadi.

Sun'iy intellekt asosidagi diagnostika tizimlari shifokorlarga yordamchi vosita sifatida xizmat qilishi mumkin. Bunday tizimlar bemor shikoyatlarini tahlil qilib, ehtimoliy kasalliklar ro'yxatini taqdim etishi yoki diagnostika jarayonida qo'shimcha ma'lumotlar berishi mumkin. Natijada tibbiy xizmat ko'rsatish jarayonining samaradorligi oshadi va bemorlar uchun tezkor maslahat tizimlari yaratiladi. Lekin, o'zbek tilida tibbiy matnlarni avtomatik tahlil qilish bo'yicha tadqiqotlar hali boshlang'ich bosqichda turibdi. Mavjud resurslarning cheklanganligi, tibbiy matnlar korpusining yetarli emasligi hamda lingvistik vositalarning kamligi ushbu yo'nalishda tadqiqot olib borishni murakkablashtiradi. Shu sababdan, o'zbek tilida tibbiy matnlarni qayta ishlash va simptomlarni aniqlashga qaratilgan ilmiy tadqiqotlar olib borish muhim ahamiyat kasb etadi. Ushbu tadqiqotning maqsadi o'zbek tilidagi bemor shikoyatlari asosida simptomlarni aniqlash va ular yordamida kasalliklarni tasniflashga mo'ljallangan modelni ishlab chiqishdan iborat. Tadqiqot jarayonida tabiiy tilni qayta ishlash usullari, matnlarni oldindan qayta ishlash texnologiyalari hamda mashinaviy

o'rganish algoritmlaridan foydalanildi. Taklif etilgan yondashuv o'zbek tilidagi tibbiy matnlarni avtomatik tahlil qilish, simptomlarni aniqlash va kasalliklarni tasniflash jarayonini samarali tashkil etishga xizmat qiladi. Ushbu tadqiqot natijalari o'zbek tilida tibbiy matnlarni qayta ishlashga mo'ljallangan intellektual tizimlarni yaratishda muhim ilmiy asos bo'lib xizmat qilishi, ishlab chiqilgan model tibbiy maslahat tizimlari, diagnostika yordamchi dasturlari va elektron tibbiy yozuvlarni tahlil qilish tizimlarini yaratishda qo'llanilishi mumkin.

Adabiyotlar tahlili. Tibbiy matnlarni avtomatik tahlil qilish va bemor simptomlari asosida kasalliklarni aniqlash masalasi tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasining muhim yo'nalishlaridan biri hisoblanadi. NLP texnologiyalari inson tilidagi matnlarni kompyuter yordamida tahlil qilish, ulardan muhim ma'lumotlarni ajratib olish hamda turli intellektual tizimlarni yaratishda keng qo'llaniladi. Ushbu jarayonlar NLPning asosiy nazariy fundamentlariga tayanadi [1:25–80]. Tabiiy tilni qayta ishlash sohasining nazariy asoslari Daniel Jurafsky va James H. Martin tomonidan yaratilgan *Speech and Language Processing* asarida batafsil yoritilgan bo'lib, unda tokenizatsiya, sintaktik va semantik tahlil hamda matn klassifikatsiyasi kabi jarayonlar keng bayon etilgan [1:150–300]. Shuningdek, matnni qayta ishlashning asosiy bosqichlari va lingvistik tahlil usullari ushbu manbada chuqur o'rganilgan. Statistik NLP usullari esa Christopher D. Manning va Hinrich Schütze tomonidan yozilgan *Foundations of Statistical Natural Language Processing* asarida batafsil bayon qilingan bo'lib, unda matnlarni ehtimollik nazariyasi asosida tahlil qilish va klassifikatsiya qilish metodlari keng yoritilgan. Xususan, ushbu asarda til modellari (n-gram), Bayes klassifikatorlari, yashirin Markov modellari (Hidden Markov Models), hamda maksimal ehtimollik (Maximum Likelihood Estimation) va smoothing usullari kabi muhim statistik yondashuvlar tushuntiriladi [2:86–340]. Mazkur metodlar matndagi so'zlar va ularning ketma-ketligi o'rtasidagi ehtimollik bog'liqliklarini aniqlash

orqali matnni tahlil qilish va bashorat qilish imkonini beradi. So‘nggi yillarda chuqur o‘rganish (deep learning) texnologiyalarining rivojlanishi natijasida transformer arxitekturasiga asoslangan modellar keng qo‘llanila boshladi. Jumladan, Devlin va boshqalar tomonidan ishlab chiqilgan BERT modeli matnni ikki tomonlama kontekst asosida tahlil qilish imkonini beradi va ko‘plab NLP vazifalarida yuqori natijalar ko‘rsatadi [3:4171–4186]. Tibbiy matnlarni tahlil qilish uchun BERT modelining maxsus variantlari ham ishlab chiqilgan. Masalan, Jinhyuk Lee va boshqalar tomonidan taklif etilgan BioBERT modeli biologik va klinik matnlarni qayta ishlash uchun mo‘ljallangan bo‘lib, simptomlarni aniqlash va kasalliklarni tasniflashda yuqori samaradorlikka ega ekanligi aniqlangan [4:835–846]. Klinik yozuvlarni tahlil qilish uchun ClinicalBERT modeli ishlab chiqilgan bo‘lib, u bemor shikoyatlari va shifokor yozuvlarini semantik jihatdan tahlil qilishda qo‘llaniladi hamda Named Entity Recognition vazifalarida samarali natijalar beradi [5:176–182]. Klinik matnlardan tibbiy ma’lumotlarni ajratib olish uchun Med7 modeli ham keng qo‘llaniladi. Ushbu model simptomlar, kasallik nomlari va dori vositalarini aniqlashga mo‘ljallangan bo‘lib, tibbiy Named Entity Recognition (NER) vazifalarida samarali hisoblanadi [6]. Bemor simptomlari asosida kasalliklarni aniqlashda mashinaviy o‘rganish algoritmlaridan foydalanish bo‘yicha ham bir qator tadqiqotlar mavjud. Ushbu tadqiqotlarda logistic regression, support vector machine (SVM) va neyron tarmoqlar kabi klassifikatsiya algoritmlari qo‘llanilib, simptomlar asosida kasalliklarni aniqlash mumkinligi isbotlangan [7:200–350]. Tibbiy matnlarni tahlil qilishda simptomlarni aniqlash muhim bosqich hisoblanadi. Buning uchun Named Entity Recognition (NER) texnologiyasidan foydalaniladi. NER tizimlari matndan muhim obyektlarni, jumladan simptomlar, kasallik nomlari va dori vositalarini avtomatik ravishda aniqlash imkonini beradi [8:350–400]

Ma’lumotlar to‘plamini shakllantirish. Tadqiqot uchun o‘zbek tilidagi bemor shikoyatlari va tibbiy matnlardan iborat ma’lumotlar to‘plami (dataset)

shakllantirildi. Dataset turli manbalardan, jumladan tibbiy forumlar, bemor murojaatlari va tibbiy maslahat platformalaridan olingan matnlar asosida tuzildi. Yig'ilgan matnlar tahlil qilinib, ulardan simptomlar va ehtimoliy kasalliklar aniqlanib, maxsus belgilash (annotatsiya) jarayoni amalga oshirildi. Har bir matn uchun mavjud simptomlar belgilandi va ular asosida ehtimoliy kasalliklar aniqlash uchun ma'lumotlar bazasi shakllantirildi.

Matnlarni oldindan qayta ishlash. Tibbiy matnlarni avtomatik tahlil qilishdan oldin ularni oldindan qayta ishlash jarayoni amalga oshirildi. Bu bosqich matnlarni standart ko'rinishga keltirish va tahlil qilishni osonlashtirish uchun muhim hisoblanadi. Matnlarni qayta ishlash quyidagi bosqichlarni o'z ichiga oldi:

- **Tokenizatsiya.** Matnni alohida so'z birliklariga ajratish
 - **Stop-wordlarni olib tashlash.** Semantik ahamiyatga ega bo'lmagan so'zlarni chiqarib tashlash
 - **Lemmatizatsiya.** So'zlarni boshlang'ich shaklga keltirish
 - **Normalizatsiya.** Turli yozuv shakllarini yagona standartga keltirish
- Mazkur jarayonlar natijasida matnlar keyingi bosqich – simptomlarni aniqlash uchun tayyor holga keltirildi.

Simptomlarni aniqlash. Matnlardan simptomlarni aniqlash uchun tabiiy tilni qayta ishlash texnologiyalaridan foydalanildi. Bunda Named Entity Recognition (NER) usulidan foydalanilib, matnda uchraydigan simptomlar avtomatik ravishda aniqlanadi. Simptomlarni aniqlash jarayonida maxsus tuzilgan simptomlar lug'ati ham qo'llanildi. Ushbu lug'atga tibbiyotda eng ko'p uchraydigan simptomlar va ularning turli variantlari kiritildi. Misol uchun:

“bosh og'rig'i”, “boshim og'riyapti”, “bosh qattiq og'riyapti”

kabi ifodalar bir xil simptom sifatida aniqlanadi.

Kasalliklarni tasniflash modeli. Simptomlar aniqlangandan so'ng ularning asosida kasalliklarni tasniflash jarayoni amalga oshirildi. Buning uchun mashinaviy

o'rganish algoritmlaridan foydalanildi. Tadqiqot davomida quyidagi klassifikatsiya algoritmlari sinovdan o'tkazildi:

- Logistic Regression
- Support Vector Machine (SVM)
- Neyron tarmoqlar (Neural Networks)

Mazkur algoritmlar yordamida simptomlar to'plami kirish ma'lumotlari sifatida qabul qilinib, ehtimoliy kasallik aniqlanadi.

Tajriba natijalari va tahlil. Taklif etilgan model samaradorligini baholash maqsadida o'zbek tilidagi bemor shikoyatlaridan iborat ma'lumotlar to'plami asosida tajribalar o'tkazildi. Tadqiqot jarayonida simptomlarni aniqlash va ular asosida kasalliklarni tasniflash uchun bir nechta mashinaviy o'rganish algoritmlari sinovdan o'tkazildi. Eksperiment davomida dataset ikki qismga ajratildi: o'qitish (training) va sinov (test) ma'lumotlari. O'qitish to'plami modelni tayyorlash uchun ishlatildi, sinov to'plami esa modelning samaradorligini baholash uchun qo'llanildi.

Model samaradorligini baholash uchun aniqlik (accuracy), aniqlash darajasi (precision), qamrov (recall) va F1 ko'rsatkichlaridan foydalanildi. Ushbu metrikalar klassifikatsiya modelining qanchalik samarali ishlashini baholash imkonini beradi. Tadqiqot davomida logistic regression, support vector machine (SVM) va neyron tarmoqlar asosidagi klassifikatsiya modellarining natijalari taqqoslab ko'rildi.

Tajriba natijalari quyidagi jadvalda keltirilgan:

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.78	0.76	0.74	0.75
SVM	0.82	0.81	0.80	0.80
Neural Network	0.87	0.86	0.85	0.85

Jadval natijalaridan ko'rinib turibdiki, neyron tarmoqlar asosidagi model boshqa algoritmlarga nisbatan yuqoriroq natijalar ko'rsatdi. Bu esa chuqur o'rganish usullarining tibbiy matnlarni tahlil qilishda samarali ekanligini ko'rsatadi. Shuningdek, tajribalar davomida simptomlarni aniqlash bosqichi modelning

umumiy natijasiga katta ta'sir ko'rsatishi aniqlandi. Simptomlarni aniqlash jarayonida matnlarni to'g'ri normalizatsiya qilish va sinonim ifodalarni bir xil shaklga keltirish model samaradorligini oshirishga yordam berdi.

Tahlil natijalari shuni ko'rsatdiki, bemor shikoyatlaridagi simptomlar asosida kasalliklarni aniqlash mumkin va bu jarayonda tabiiy tilni qayta ishlash texnologiyalaridan foydalanish samarali natijalar beradi. Taklif etilgan model tibbiy matnlarni avtomatik tahlil qilish va diagnostika tizimlarini yaratishda qo'llanilishi mumkin.

Xulosa. Mazkur tadqiqot natijalari o'zbek tilidagi bemor shikoyatlari asosida simptomlarni aniqlash va ular yordamida kasalliklarni tasniflash mumkinligini tasdiqladi. Tabiiy tilni qayta ishlash (NLP) texnologiyalaridan foydalanish, xususan matnlarni oldindan qayta ishlash bosqichlari (tokenizatsiya, lemmatizatsiya, normalizatsiya) hamda Named Entity Recognition (NER) usullarini qo'llash simptomlarni aniqlash aniqligini sezilarli darajada oshirishi aniqlandi. Tajriba natijalariga ko'ra, logistic regression, SVM va neyron tarmoqlar orasida aynan neyron tarmoqlar eng yuqori samaradorlikni ko'rsatdi. Bu esa chuqur o'rganish usullarining o'zbek tilidagi tibbiy matnlarni tahlil qilishda yuqori salohiyatga ega ekanligini ko'rsatadi. Shu bilan birga, simptomlarni aniqlash bosqichida sinonim ifodalarni yagona shaklga keltirish, matnlarni to'g'ri normalizatsiya qilish va kontekstni hisobga olish model natijalariga bevosita ta'sir qiluvchi muhim omillar ekanligi qayd etildi. Tadqiqotning amaliy ahamiyati shundaki, ishlab chiqilgan model tibbiy matnlarni avtomatik tahlil qilish, bemor shikoyatlaridan muhim klinik ma'lumotlarni ajratib olish hamda ehtimoliy kasalliklarni aniqlash imkonini beradi. Ushbu yondashuv tibbiy maslahat tizimlari, elektron tibbiy yozuvlarni qayta ishlash platformalari va shifokorlar uchun yordamchi diagnostika vositalarini yaratishda samarali qo'llanilishi mumkin. Natijada diagnostika jarayoni tezlashadi, inson omiliga bog'liq xatolar kamayadi va tibbiy xizmat sifati oshadi. Bundan tashqari,

tadqiqot o'zbek tilida tibbiy matnlarni qayta ishlash bo'yicha ilmiy izlanishlar uchun muhim asos bo'lib xizmat qiladi. Kelgusida ushbu yo'nalishda katta hajmdagi annotatsiyalangan datasetlar yaratish, transformer asosidagi modellarni o'zbek tiliga moslashtirish, shuningdek, multimodal ma'lumotlar asosida yanada mukammal diagnostika tizimlarini ishlab chiqish istiqbolli vazifalardan biri hisoblanadi.

Foydalanilgan adabiyotlar ro'yxati

1. Alsentzer E., Murphy J. R., Boag W. et al. Publicly Available Clinical BERT Embeddings // Proceedings of Clinical NLP Workshop (ACL). – 2019. – P. 176–182.
2. Bishop C. M. Pattern Recognition and Machine Learning. – New York: Springer, 2006. – 738 p.
3. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186.
4. Jurafsky D., Martin J. H. Speech and Language Processing. – 3rd ed. – Pearson, 2023. – 1024 p.
5. Kormilitzin A., Vaci N., Liu Q., Nevado-Holgado A. Med7: A transferable clinical natural language processing model for electronic health records // arXiv preprint. – 2020. – arXiv:2003.01271.
6. Lee J., Yoon W., Kim S. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining // Bioinformatics. – 2020. – Vol. 36, No. 4. – P. 835–846.
7. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. – Cambridge: MIT Press, 1999. – 680 p.