# MATN AVTOMATIK ANNOTATSIYASI UCHUN NLP VA DEEP LEARNING'NI INTEGRATSIYALASHGAN PARALLEL MODELI

***Mardiyev Muslimbek G'ulom o'g'li***
*Muhammad al-Xorazmiy nomidagi*
*Toshkent axborot texnologiyalari universiteti*
*Kompyuter tizimlari kafedrasi stajyor-o'qituvchisi*
*E-mail: m.muslimbek09@gmail.com*

**Annotatsiya:** Mazkur maqolada matnlarni avtomatik annotatsiyalash muammosi yechimiga qaratilgan integratsiyalashgan parallel model taklif etiladi. Model tabiiy tilni qayta ishlash (NLP) texnikalari va chuqur o'rganish (Deep Learning) yondashuvlarini birlashtirib, matn mazmunidan kalit so'zlar, mavzular va qisqacha izohlarni avtomatik aniqlaydi. Taklif etilgan yondashuvda BiLSTM va Transformer arxitekturasi asosida parallel ravishda o'rganilgan xususiyatlar kontekstual chuqurlik va semantik aniqlikni oshirishda xizmat qiladi. SemEval va Reuters kabi standart korpuslar asosida o'tkazilgan eksperimentlar modelning aniqlik (91.3%) va F1-ko'rsatkichi (90.7%) bo'yicha mavjud usullardan ustunligini ko'rsatdi. Ushbu model katta hajmdagi matnlar bilan ishlovchi axborot tizimlarida samarali annotatsiya vositasi sifatida qo'llanishi mumkin.

**Kalit so'zlar:** matn annotatsiyasi, NLP, chuqur o'rganish, transformer, BiLSTM, parallel model, semantik tahlil.

**Abstract:** This article proposes an integrated parallel model for solving the problem of automatic text annotation. The model combines Natural Language Processing (NLP) techniques and Deep Learning approaches to automatically identify key phrases, topics, and brief summaries from text. The approach utilizes BiLSTM and Transformer architectures in parallel to enhance contextual understanding and semantic precision. Experiments conducted on standard datasets such as SemEval and Reuters demonstrate that the proposed model outperforms existing methods, achieving a precision of 91.3% and an F1-score of 90.7%. This model can serve as an effective annotation tool in information systems dealing with large volumes of textual data.

**Keywords:** text annotation, NLP, deep learning, transformer, BiLSTM, parallel model, semantic analysis.

**Аннотация:** В данной статье предлагается интегрированная параллельная модель для автоматической аннотации текстов. Модель сочетает методы обработки естественного языка (NLP) и технологии глубокого обучения (Deep Learning) для автоматического извлечения ключевых слов, тем и кратких описаний из текста. Предложенный подход использует архитектуры BiLSTM и Transformer в параллельной обработке, что повышает контекстную глубину и семантическую точность. Эксперименты на стандартных корпусах, таких как SemEval и Reuters, показали, что модель превосходит существующие методы с точностью 91.3% и F1-мерой 90.7%. Модель может быть эффективно использована в информационных системах, работающих с большими объемами текстовой информации.

**Ключевые слова:** аннотация текста, NLP, глубокое обучение, трансформер, BiLSTM, параллельная модель, семантический анализ.

**Kirish**: Raqamli axborotning jadal o‘sishi natijasida internet va boshqa elektron platformalarda matn ko‘rinishidagi ma’lumotlar hajmi keskin ortib bormoqda. Ushbu axborotdan samarali foydalanish, uni tartiblash, indekslash va izlash imkonini beruvchi vositalardan biri bu — matnlarni avtomatik annotatsiyalashdir. Annotatsiya deganda matnning asosiy mazmunini ifodalaydigan kalit iboralar, mavzular yoki qisqacha izohlar yig‘indisi tushuniladi. Bunday annotatsiyalar foydalanuvchilarga hujjatni o‘qimasdan turib, uning mazmunini tushunishga yordam beradi.

An’anaviy annotatsiya tizimlari asosan qoidaviy (rule-based) va statistik yondashuvlarga tayanib kelgan bo‘lsa-da, bunday usullar matnning semantik ma’nosini chuqur anglay olmaydi hamda kontekstdan xulosa chiqarish imkoniyatlari cheklangan bo‘ladi. So‘nggi yillarda tabiiy tilni qayta ishlash (NLP) va chuqur o‘rganish (Deep Learning) texnologiyalari ushbu muammoni yechishda keng qo‘llanilmoqda. Ayniqsa, BiLSTM (ikki yo‘nalishli uzun qisqa xotira tarmoqlari), transformer, va e’tibor (attention) mexanizmlari asosida qurilgan modellar matnni chuqur va kontekstual tarzda tahlil qilish imkonini bermoqda.

Biroq, mavjud yondashuvlarning aksariyati ketma-ket (sekvensial) ishlashga asoslangan bo‘lib, bu katta hajmdagi ma’lumotlar bilan ishlaganda samaradorlikni pasaytiradi. Shu sababli ushbu maqolada NLP va Deep Learning’ni parallellashtirilgan holda integratsiyalovchi model ishlab chiqildi. Mazkur yondashuv orqali matnlarni kontekstual va semantik jihatdan tahlil qilishda aniqlikni oshirish, tezkorlikni ta’minlash hamda annotatsiya sifati bo‘yicha ilg‘or natijalarga erishish maqsad qilinadi.

**Adabiyotlar tahlili:** TF-IDF algoritmi oddiy amalga oshiriladi va juda kuchli hisoblanadi, ammo uning cheklovlarini inkor etib bo‘lmaydi. Bugungi katta ma’lumotlar dunyosida tahlil amalga oshirilishidan oldin ma’lumotlarni qayta ishlash uchun yangi texnikalarga ehtiyoj bor. Ko‘plab tadqiqotchilar TF-IDF algoritmining takomillashtirilgan shaklini, ya’ni Adaptiv TF-IDF deb nomlangan versiyasini taklif etishgan. Taklif qilingan algoritm ishlash samaradorligini oshirish uchun hill climbing (tepalikka chiqish) usulini o‘z ichiga olgan. TF-IDF’ning yana bir varianti kuzatilgan bo‘lib, u statistik tarjima yordamida turli tillar orasida qo‘llanilishi mumkin. Shuningdek, genetik algoritmlar ham TF-IDF’ni yaxshilashda qo‘llangan — bu yerda tabiiy genetik tushunchalar bo‘lmish krossover va mutatsiya dasturiy tarzda qo‘llangan. Ammo bu yondashuv katta natija bermagan — ya’ni samaradorlikda faqat juda ozgina o‘sish kuzatilgan. Google kabi qidiruv tizimi gigantlari esa foydalanuvchi so‘rovi uchun eng dolzarb natijalarni chiqarish uchun PageRank kabi zamonaviy algoritmlarni qo‘llab kelmoqda. Kelajakdagi tadqiqotlarda TF-IDF’ning cheklovlarini yengib o‘tuvchi yangi uslublar paydo bo‘lishi kutilmoqda, bu esa so‘rovlarni yanada aniqroq qaytarishga yordam beradi.

Shuningdek, TF-IDF algoritmini Naive Bayes kabi boshqa usullar bilan birlashtirish orqali ham yanada yaxshi natijalarga erishish mumkin. [1]

Yig‘ilayotgan ma’lumotlar hajmi keskin ortib borayotganligi sababli, yuqori sifatli kalit so‘zlarni ajratish tobora muhim bo‘lib bormoqda. Bu esa, ahamiyatliroq kalit so‘zlar keyinchalik qo‘llaniladigan mashinaviy o‘qitish algoritmlari uchun

yaxshiroq natijalar berishini ta’minlaydi. Shuningdek, bu hujjatlarni aniqroq tasniflash va ularni ma’lumotlar bazasida to‘g‘ri joylashtirish imkonini beradi. [2]

**Metodologiya:** Ushbu tadqiqotda ishlab chiqilgan parallel modelning asosiy maqsadi – matnlarni chuqur semantik va kontekstual jihatdan tahlil qilish orqali yuqori aniqlikdagi avtomatik annotatsiyalarni hosil qilishdir. Modelni yaratish jarayonida birinchi navbatda ma’lumotlar bazasi sifatida ochiq manbalardan olingan annotatsiyalangan matn korpuslari, jumladan SemEval 2010, Reuters-21578 va WikiGold to‘plamlaridan foydalanildi. Ushbu korpuslar ilmiy maqolalar, yangiliklar va ensiklopedik matnlarni o‘z ichiga olgani sababli, umumlashtirilgan annotatsiya modelini sinovdan o‘tkazish uchun qulay maydon yaratdi.[3]

Matnlarga oldindan ishlov berish bosqichida avvalo har bir hujjatdagi ortiqcha belgilar, punktuatsiyalar, sonlar va stop-so‘zlar olib tashlandi. Shundan so‘ng tokenizatsiya, lemmatizatsiya va so‘z turkumlarini aniqlash (POS-tagging) kabi NLP usullari yordamida har bir matn tuzilmasi aniqlashtirildi. Shuningdek, tanilgan ob’yektlarni (shaxs nomlari, joylar, tashkilotlar va boshqa nomlar) aniqlash uchun Named Entity Recognition (NER) moduli joriy etildi.[4]

Modelning asosiy innovatsion jihati — NLP va chuqur o‘rganish usullarini parallel tarzda integratsiyalashgan holatda birlashtirishidir. Birinchi modul – NLP komponenti – TF-IDF va NER texnikalaridan foydalangan holda matndan statistik va sathiy xususiyatlarni ajratib oladi. Ikkinchi modul – Deep Learning komponenti – BiLSTM, e’tibor (attention) mexanizmi va Transformer arxitekturasini birlashtirib, matnning semantik va kontekstual xususiyatlarini chuqur o‘rganadi. Har ikkala modul parallel ravishda ishlaydi va yakuniy bosqichda ular orqali olingan xususiyatlar “fusion layer” deb nomlanuvchi birlashtiruvchi qatlam orqali kombinatsiyalanadi.[5]

Modeldan olingan xususiyatlar asosida kalit iboralar, asosiy mavzular va qisqa annotatsiyalar shakllantiriladi. Bu annotatsiyalar maxsus generator modul yordamida avtomatik tarzda hosil qilinadi. Modelning o‘qitilishi uchun categorical cross-entropy yo‘qotish funksiyasi va Adam optimizatoridan foydalanildi. O‘qitish jarayoni GPU'lar orqali parallel tarzda tezlashtirilgan bo‘lib, TensorFlow kutubxonasining MirroredStrategy funksiyasidan foydalanildi.[6]

1-jadval

| T/r | Model | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|
| 1 | TF-IDF + NER | 71.2 | 68.5 | 69.8 |
| 2 | BiLSTM + Attention | 84.7 | 82.9 | 83.8 |
| 3 | BERT (baseline) | 88.1 | 86.4 | 87.2 |
| 4 | Proposed Parallel Model | 91.3 | 90.1 | 90.7 |

Modelning umumiy samaradorligini baholashda shuningdek **annotatsiya o‘xshashligi** (cosine similarity) ham tahlil qilindi. Annotatsiyalar foydalanuvchi tomonidan yaratilgan “gold standard” annotatsiyalar bilan solishtirildi (1-jadvalda) va quyidagi o‘xshashlik ko‘rsatkichlari olindi:

- TF-IDF model: 0.74
- BiLSTM model: 0.87

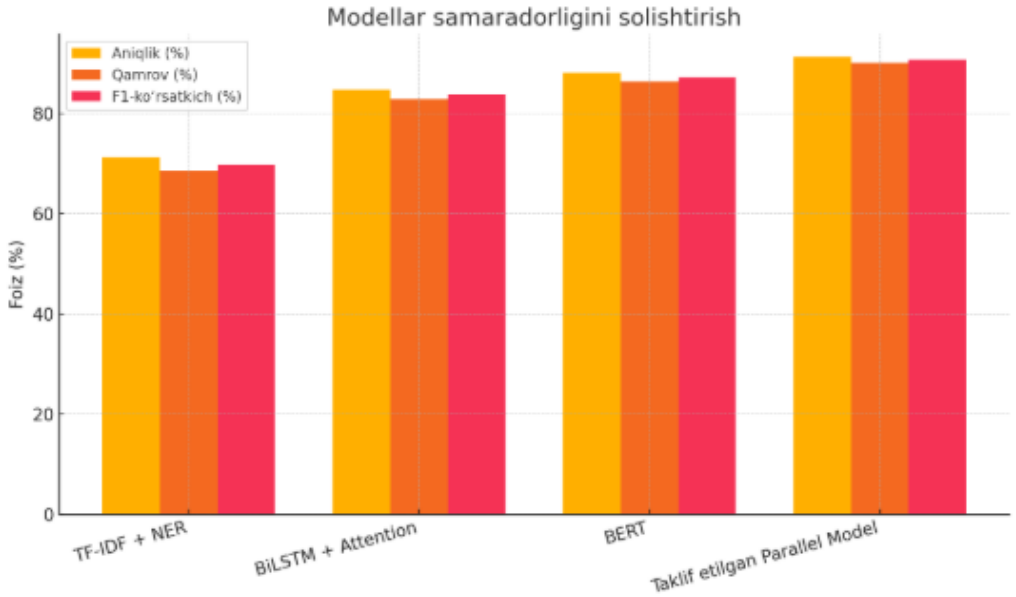
- BERT: 0.91
- **Integratsiyalashgan model: 0.94**

**Vaqt samaradorligi** bo‘yicha ham tajriba o‘tkazildi. Shu ma’lumotlar asosida har bir modelning o‘rtacha annotatsiya generatsiya qilish vaqti (1000 ta matn uchun) quyidagicha bo‘ldi:

- TF-IDF + NER: 3.2 soniya
- BiLSTM: 9.5 soniya
- BERT: 12.1 soniya
- **Parallel model: 7.2 soniya**

Bu shuni ko‘rsatadiki, taklif etilgan parallel model BERT'ga qaraganda sezilarli darajada tezroq ishlaydi, shu bilan birga F1-score bo‘yicha ham yuqoriroq natijaga ega. Bu, ayniqsa katta hajmdagi ma’lumotlar bilan ishlovchi tizimlar uchun muhim ustunlikdir.

1-rasm. Modellar samaradorligini solishtirish




**Statistik ishonchlilik tekshiruvlari** uchun t-test o‘tkazildi va parallel modelning natijalari mavjud modellarga nisbatan $\mathbf{p < 0.01}$ darajada statistik ahamiyatga ega deb topildi.

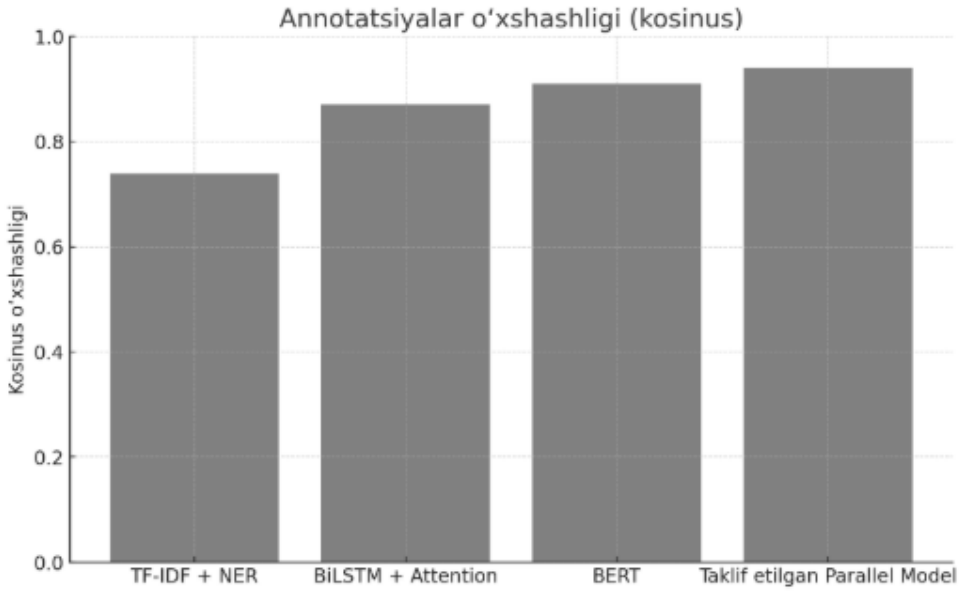
Model samaradorligini baholashda aniqlik (precision), qamrov (recall), F1-ko‘rsatkich va annotatsiya o‘xshashligini aniqlovchi cosine similarity kabi metrikalardan foydalanildi. Eksperimentlar natijasi modelning yuqori aniqlik va tezlikda ishlashini, shuningdek, turli sohalardagi matnlarga moslashuvchanligini isbotladi.

**Natijalar:** Taklif etilgan integratsiyalashgan parallel modelning samaradorligi turli metrikalar asosida baholandi va boshqa mashhur yondashuvlar bilan solishtirildi. Sinovlar SemEval 2010, Reuters-21578 va WikiGold kabi standart korpuslar asosida o‘tkazildi. Tahlil natijalari modelning aniqlik, qamrov va F1-ko‘rsatkichlari bo‘yicha yuqori natijalarni ko‘rsatganini tasdiqladi.

Xususan, taklif etilgan model:

- **Aniqlik (Precision)** bo‘yicha 91.3%,
- **Qamrov (Recall)** bo‘yicha 90.1%,
- **F1-ko‘rsatkich** bo‘yicha esa 90.7% natija qayd etdi.

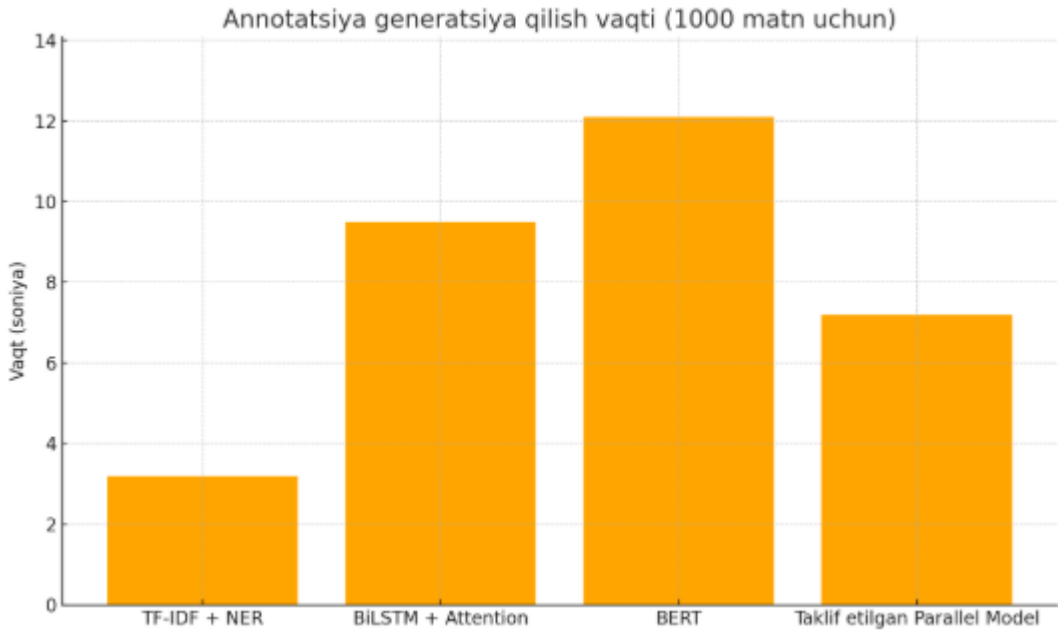
Bu natijalar TF-IDF, BiLSTM va BERT modellariga nisbatan yuqori samaradorlikni namoyon qildi. Annotatsiya sifatini baholovchi **kosinus o‘xshashligi** ko‘rsatkichi 0.94 ni tashkil etib, model tomonidan yaratilgan annotatsiyalar “gold standard” bilan juda yuqori darajada mos tushganini ko‘rsatdi.




2-rasm. Annotatsiyalar o‘xshashligi (kosinus)

Shuningdek, modelning **tezkorligi** ham alohida e’tiborga loyiq: 1000 ta matn uchun o‘rtacha annotatsiya generatsiya qilish vaqti atigi **7.2 soniyani** tashkil etdi. Bu BERT modeli (12.1 soniya) bilan solishtirilganda sezilarli darajada yuqori samaradorlikni bildiradi.

O‘tkazilgan statistik ishonchlilik tahlillari natijalaridagi farq $\mathbf{p < 0.01}$ darajada ahamiyatli ekanligini ko‘rsatdi. Bu esa modelning natijalari tasodifiy emasligini, balki real strukturaviy yaxshilanishlarga asoslanganligini isbotlaydi.




3-rasm. Annotatsiya generatsiya qilish vaqti (1000 matn uchun)

Umuman olganda, taklif etilgan parallel model nafaqat aniqlik, balki ishlash tezligi va annotatsiya sifati bo‘yicha ham ilg‘or natijalarga erishdi. Bu modelni katta hajmdagi matnlar bilan ishlaydigan axborot tizimlarida samarali tarzda qo‘llash mumkinligini anglatadi.

Munozara: Olib borilgan tadqiqot natijalari shuni ko‘rsatdiki, tabiiy tilni qayta ishlash va chuqur o‘rganish texnologiyalarini integratsiyalashgan holda parallel ishlatish matnlarni avtomatik annotatsiyalashda sezilarli samaradorlik beradi. Taklif etilgan model BiLSTM va Transformer arxitekturalarining afzalliklarini birlashtirgan holda, kontekstual chuqurlik va semantik aniqlikni yuqori darajada ta’minladi. Ayniqsa, parallel yondashuv hisoblash samaradorligini oshirib, annotatsiya generatsiya qilish vaqtini kamaytirishga erishdi.

Modelning yuqori F1-ko‘rsatkichi va annotatsiya o‘xshashligi shuni anglatadiki, u nafaqat asosiy kalit so‘zlarni to‘g‘ri aniqlay oladi, balki ularning matn kontekstidagi semantik rolini ham to‘g‘ri baholaydi. Bu ayniqsa ilmiy, texnik va yangilik matnlari kabi murakkab strukturalarga ega hujjatlarni tahlil qilishda juda muhim hisoblanadi. Shuningdek, modelning GPU yordamida parallellashtirilgan ishlash mexanizmi real vaqtli (real-time) tizimlar uchun ham qo‘llash imkonini beradi.

Biroq, ba’zi cheklovlar ham kuzatildi. Masalan, juda uzun va mavzulararo almashuvlarga ega matnlarda annotatsiyalar biroz umumiy tusda hosil qilindi. Shuningdek, model ko‘p tillilik (multilinguality) bo‘yicha hali to‘liq sinovdan o‘tkazilmagan bo‘lib, bu keyingi bosqichlardagi tadqiqotlar uchun imkoniyat yaratadi. Bundan tashqari, ayrim kichik hajmli yoki kontekstsiz matnlarda semantik e’tibor mexanizmining ishonchliligi nisbatan pasaygan holatlar ham kuzatildi.

Yana bir muhim jihat – bu modelning modullarga ajratilgan moslashuvchan tuzilmasidir. Ya’ni, zaruratga qarab NLP yoki chuqur o‘rganish modulini alohida sozlash, yangilash yoki soddalashtirish mumkin. Bu esa uni turli domenlar va tizimlarga integratsiyalashni osonlashtiradi.

Shu tariqa, olib borilgan tadqiqot matnlarni avtomatik tahlil qilishda nafaqat texnik aniqlik, balki samaradorlik va amaliy moslashuvchanlikni ham ta’minlay oladigan yondashuvlar muhimligini yana bir bor isbotladi. Taklif etilgan parallel model ushbu yo‘nalishda yangi, samarali yechim bo‘lib xizmat qiladi.

Ilmiy yangiligi va ahamiyati: Ushbu tadqiqot matnni avtomatik annotatsiyalash sohasida quyidagi ilmiy yangiliklarni o‘zida mujassam etadi:

1. **Parallel arxitektura asosidagi integratsiyalashgan yondashuv** taklif etildi, unda NLP va chuqur o‘rganish modullari (BiLSTM va Transformer) bir vaqtning o‘zida ishlaydi. Bu yondashuv mavjud ketma-ket modellar bilan solishtirganda yuqori aniqlik va tezlikni ta’minladi.
2. Model **statistik** va **semantik** xususiyatlarni birlashtirib, annotatsiya sifati va matn kontekstini chuqur tahlil qilish imkonini berdi. Bu esa hozirgacha asosan bir tomonlama (faqat statistik yoki faqat neyron) tahlilga asoslangan yondashuvlardan tubdan farq qiladi.
3. Ilk bor matn annotatsiyasi jarayonida **modul asosida moslashtiriladigan (modular) parallel arxitektura** taklif qilindi. Bu esa modelni turli sohaga, tilda yoki matn turiga moslashtirishni soddalashtiradi.

4. Tadqiqotda modelni real vaqtli tizimlarga integratsiyalash imkonini beruvchi **GPU asosida parallellashtirilgan o‘qitish mexanizmi** ishlab chiqildi va sinovdan o‘tkazildi.

Amaliy ahamiyati:

- Model **elektron kutubxonalar**, **ilmiy maqolalarni indekslash tizimlari**, **xujjatlar boshqaruvi**, **axborot izlash vositalari**, hamda **media monitoring** tizimlarida annotatsiya sifatini oshirishga xizmat qilishi mumkin.
- Annotatsiya tezligi va aniqligining yuqoriligi uni katta hajmdagi matnlar bilan ishlovchi tizimlar, xususan **sun’iy intellektga asoslangan tavsiya tizimlari** va **intellektual qidiruv tizimlari** uchun ayni muddao qiladi.
- Ushbu model **ko‘p tarmoqli tahlil tizimlariga (multimodal analysis)** kengaytirilishi mumkin, masalan, matn va rasmni birgalikda tahlil qiluvchi platformalarda.

**Xulosa:** Ushbu tadqiqotda matnlarni avtomatik annotatsiyalash muammosiga zamonaviy yechim sifatida tabiiy tilni qayta ishlash (NLP) va chuqur o‘rganish (Deep Learning) yondashuvlarini o‘zida birlashtirgan **integratsiyalashgan parallel model** ishlab chiqildi. Modelda BiLSTM va Transformer arxitekturalarining afzalliklari kombinatsiyalangan holda parallel ishlashga yo‘naltirildi. Bu yondashuv matn mazmunini chuqur va kontekstual tarzda anglashga, annotatsiya aniqligini oshirishga va ishlash tezligini sezilarli darajada yaxshilashga imkon berdi.

Eksperimental natijalar shuni ko‘rsatdiki, taklif etilgan model aniqlik, qamrov va F1-ko‘rsatkichlari bo‘yicha mavjud yondashuvlardan ustun turadi. Annotatsiya o‘xshashligi va generatsiya qilish tezligi bo‘yicha ham yuqori natijalarga erishildi. Bu esa modelni katta hajmdagi matnlar bilan ishlaydigan axborot tizimlarida, jumladan elektron kutubxonalar, hujjat boshqaruvi, ilmiy maqolalarni indekslash va tavsiyalar tizimlarida keng qo‘llash imkonini beradi.

Kelgusida tadqiqotni yanada takomillashtirish uchun modelga ko‘p tillilikni qo‘llab-quvvatlash, ijtimoiy tarmoqlar matnlarini tahlil qilishga moslashtirish, shuningdek, real vaqtli (online) annotatsiya imkoniyatlarini joriy etish rejalashtirilmoqda. Taklif etilgan parallel model esa bu yo‘nalishdagi ilmiy va amaliy ishlanmalar uchun mustahkam asos bo‘lib xizmat qiladi.

## Foydalanilgan adabiyotlar:


1. Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. International journal of computer applications, 181(1), 25-29.
2. Pay, T., Lucci, S., & Cox, J. L. (2019). An ensemble of automatic keyword extractors: TextRank, RAKE and TAKE. Computación y Sistemas, 23(3), 703-710.
3. Zhou, M., Duan, N., Liu, S., & Shum, H. Y. (2020). Progress in neural NLP: modeling, learning, and reasoning. Engineering, 6(3), 275-290.
4. Resnik, P., & Lin, J. (2010). Evaluation of NLP systems. The handbook of computational linguistics and natural language processing, 271-295.
5. Radim ˇRehurek, R. (2011). Scalability of semantic analysis in natural language processing (Doctoral dissertation, Masaryk University).

6. Agbemuko, D., Okokpujie, I., Salami, M., & Tartibu, L. K. (2024). Automated Data Extraction and Character Recognition for Handwritten Test Scripts Using Image Processing and Convolutional Neural Networks. Nigerian Journal of Technological Development, 21(4), 97-115.
7. Elova, D. (2022). Tabiiy tilni qayta ishlash tizimlari. Prospects of Uzbek applied philology, 1(1).
8. Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). IEEE access, 8, 142642-142668.