NAZORAT OSTIDAGI O‘QITISH ALGORITMLARI YORDAMIDA SO‘Z MA’NOSINI ANIQLASH

Axmedova Xolisxon Ilxomovna
O‘zbekiston xalqaro islomshunoslik akademiyasi katta o‘qituvchisi, texnika fanlari falsafa doktori(PhD)
E-mail: xolisa9029@mail.ru
Tel: +998944189029
Orcid: 0000-0002-9828-1650

Annotatsiya. Tabiiy tilni qayta ishlash (NLP) sohasida semantik tahlil hanuz eng muhim va dolzarb yo‘nalishlardan biri hisoblanadi. Ushbu tahlilning asosiy vazifalaridan biri – so‘z ma’nolarini aniqlash (Word Sense Disambiguation, WSD) bo‘lib, omonim so‘zlarni avtomatik aniqlash jarayonida zamonaviy algoritmlardan foydalanish talab etiladi. Mazkur ishda ushbu muammoni yechish uchun K-yaqin qo‘shni (K-NN), Random Forest algoritmlari hamda SenseBERT modelidan foydalanildi. O‘zbek tilidagi omonim so‘zlarning turli ma’nolariga oid jumlalarni qamrab olgan maxsus dataset bazasi tuzildi va dastlabki ishlov berish bosqichlaridan o‘tkazildi. Ma’lumotlar tozalangach, Random Forest va SenseBERT algoritmlari asosida modellar o‘qitildi. Sinov bosqichida o‘qitilgan modellar o‘rtacha 92 % aniqlikka erishdi. Kelgusida natijalarni yanada yaxshilash uchun korpus hajmini kengaytirish hamda yuqori xotiraga ega hisoblash resurslarida modellarni qayta o‘qitish tavsiya etiladi.

Kalit so‘zlar: so‘z ma’nosini aniqlash, mashinali o‘qitish yondashuvlari, KNN algoritmi, Random Forest algoritmi, SenseBERT modeli, ma’lumotlar to‘plami.

Abstract. In the field of Natural Language Processing (NLP), semantic analysis is still considered one of the most important and relevant research directions. One of the main tasks of this analysis is Word Sense Disambiguation (WSD), which requires the use of modern algorithms for the automatic identification of homonymous words. In this study, the K-Nearest Neighbors (K-NN), Random Forest algorithms, and the SenseBERT model were employed to solve this problem. A specialized dataset was created that includes sentences covering different meanings of homonymous words in the Uzbek language, and the data were processed through preliminary preprocessing stages. After data cleaning, models were trained using the Random Forest and SenseBERT algorithms. During the testing phase, the trained models achieved an average accuracy of 92%. For further improvement of the results, it is recommended to expand the corpus size and retrain the models using computational resources with higher memory capacity.

ISSN 3030-3907

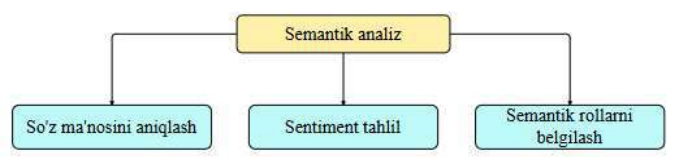
Keywords: Word Sense Disambiguation, machine learning approaches, KNN algorithm, Random Forest algorithm, SenseBERT model, dataset.

Аннотация. В области обработки естественного языка (NLP) семантический анализ по-прежнему считается одним из наиболее важных и актуальных направлений исследований. Одной из основных задач данного анализа является разрешение многозначности слов (Word Sense Disambiguation, WSD), при котором для автоматического выявления омонимичных слов требуется использование современных алгоритмов. В данной работе для решения этой задачи были использованы алгоритмы K ближайших соседей (K-NN), Random Forest, а также модель SenseBERT. Была создана специализированная база данных, охватывающая предложения с различными значениями омонимичных слов в узбекском языке, и проведены этапы предварительной обработки данных. После очистки данных модели были обучены на основе алгоритмов Random Forest и SenseBERT. На этапе тестирования обученные модели достигли средней точности 92 %. В дальнейшем для улучшения результатов рекомендуется расширить объём корпуса и повторно обучить модели на вычислительных ресурсах с большим объёмом памяти.

Ключевые слова: разрешение многозначности слов, подходы машинного обучения, алгоритм K ближайших соседей (KNN), алгоритм случайного леса (Random Forest), модель SenseBERT, набор данных.

KIRISH.

XXI asrda eng tez rivojlanayotgan yo‘nalishlardan biri bu – kompyuter lingvistikasi sanaladi. Ushbu fan tabiiy tilga ishlov berish sohasining turli vazifalarini o‘z ichiga oladi va mazkur vazifalarni turli tabiiy tillar kesimida hal qilish nihoyatda muhim ahamiyatga ega. Mashina tarjimasi (MT), savol-javob tizimlarini yaratish, matnlarni tasniflash hamda nomlangan obyektlarni aniqlash (NER) masalalari shular jumlasidandir. Har bir masalaning yechimi o‘z ichida bir nechta kichik bosqichlarni qamrab oladi. Masalan, matnlarni tahlil qilish uchta: morfologik tahlil [1; 27-38-b.], sintaktik tahlil [2; 46-58-b.] va semantik tahlil bosqichlaridan iborat. Ushbu tahlil turlari o‘ziga xos iyerarxik jarayonlarga asoslanadi. Mazkur maqolada semantik tahlil va uning asosiy tarkibiy qismlariga umumiy nuqtayi nazardan yondashuv yoritib beriladi.

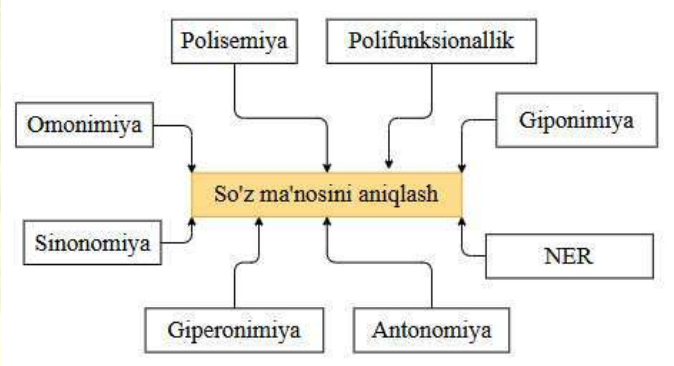


1-rasm:Semantik tahlil turlari

ISSN 3030-3907

Semantik tahlil, o‘z navbatida, uchta asosiy yo‘nalishga ajratiladi: Sentiment tahlil qilish, so‘z ma’nosini aniqlash hamda semantik rollarni belgilash. Birinchisi matnning hissiy ohangini baholashga xizmat qilsa, ikkinchisi so‘zning ma’lum bir kontekstdagi aniq ma’nosini belgilashga yordam beradi. Semantik rollarni belgilash esa gapdagi har bir so‘zning vazifasini aniqlash demakdir. So‘z ma’nosini aniqlash NLPning muhim vazifalaridan biri sifatida katta ahamiyat kasb etadi.

Tabiiy til birliklaridagi so‘zlarning to‘g‘ri ma’nosini aniqlash muammosini hal etish, asosan, turli turdagi so‘zlarni semantik jihatdan farqlash orqali amalga oshiriladi.



2-rasm: So‘z ma’nosini aniqlash elementlari

Jahonning ko‘plab tabiiy tillarida so‘z ma’nosini aniqlash muammosi chuqur o‘rganilgan va turli darajada yechim topgan. Quyida bu vazifani bajarishda keng qo‘llaniladigan asosiy yondashuvlar keltiriladi:

1. **Bilimga asoslangan yondashuvlar**
2. **Mashinali o‘qitish yondashuvlari**

2.1. Nazorat ostida o‘qitish algoritmlari

2.2. Nazoratsiz o‘qitish algoritmlari

2.3. Yarim nazorat ostida o‘qitish algoritmlari

3. **Neyron tarmoq modellar**

MAVZUGA OID ADABIYOTLAR TAHLILI.

So‘z ma’nosini aniqlash bo‘yicha turli tillarda juda ko‘p tadqiqotlar amalga oshirilgan. So‘nggi yillarda mazkur masalani yechishda zamonaviy uslublardan tobora ko‘proq foydalanilmoqda. Masalan, rus tilida omonimlarni aniqlash uchun neyron tarmoqlardan foydalanilgan [3; 56-64-b.]; qozoq [4; 4-13-b.], arab [5; 5-15-b.], hind [6; 78-83-b.], ingliz [7; 40-51-b.] hamda rus [8; 9-14-b.] tillarida ham omonimiya va polisemiya hodisalarini aniqlashda turli mashinali o‘qitish va chuqur o‘rganish usullari qo‘llangan. O‘zbek tilida ham so‘zlarning ma’nosini aniqlashga qaratilgan tadqiqotlar uchraydi.

So‘z ma’nosini aniqlash masalasi bilimga asoslangan usullar yordamida ham bajarilishi mumkin, biroq bunday yondashuvlarning aniqligi nisbatan past bo‘ladi.

ISSN 3030-3907

Ular ko‘pincha ma’lum bir soha doirasidagi noaniq so‘zlarning kontekstual ma’nosini topishda foydalaniladi [9; 305-319-b.],[2; 27-38-b],[10; 137-140-B.].

Shu bilan birga, polisemiya hodisasiga ega bo‘lgan so‘zlarning ma’nosini kontekstda farqlash uchun mashinali o‘qitish yondashuvlari yanada samaraliroq ekanligi isbotlangan [11; 2825-2830-b.]. Nazorat ostida o‘qitish algoritmlari esa tegishli semantik teglar bilan belgilangan ma’lumotlar to‘plamini talab qiladi. Bunday to‘plamdagi har bir yozuv bir qancha xususiyatlar ($f_1 \ldots f_n$) bilan ifodalanadi. Algoritmning vazifasi ana shu xususiyatlar va ularning tegishli toifalari ($k_1 \ldots k_n$) o‘rtasidagi bog‘liqlikni o‘rganib, yangi, ilgari ko‘rilmagan ma’lumotlarni ham o‘rgangan teglar asosida to‘g‘ri tasniflashga qodir bo‘lishidir.

Kontekstda so‘z ma’nosini aniqlash uchun Nazorat ostida o‘qitish usullari semantik belgilangan ma’lumotlar to‘plami yoki korpusga tayanadi. Masalan, ingliz tilidagi polisemiya hodisasiga ega bo‘lgan so‘zlarni ajratishda SemCor korpusi keng qo‘llaniladi. SemCor – Brown korpusining taxminan 234 ming so‘zdan iborat kichik qismi bo‘lib, undagi har bir gapdagi leksik birliklar WordNetdagi tegishli ma’nolari bilan belgilangan. Ushbu WordNet teglari tabiiy ravishda inson annotatorlari tomonidan qo‘lda kiritilgan [12; 349-370-b.].

TADQIQOT METODOLOGIYASI.

Bunday algoritmlar amalda qo‘llanganda, odatda ikki asosiy turdagi xususiyatlardan foydalaniladi va ular turli shakllarda kombinatsiya qilinadi: *kollokatsion xususiyatlar* va *birga uchrash xususiyatlari*.

Kollokatsion xususiyatlar deganda, maqsadli so‘zning chap yoki o‘ng tarafida ma’lum pozitsiyalarda uchraydigan aniq so‘zlar va ularning so‘z turkumlari nazarda tutiladi. Masalan:

*“Traktor matorining shovqini suruvdagi **otlarni** hurkitib yubordi”* Agar ushbu gapda maqsadli so‘z ***“ot+larni”*** bo‘lsa, unda ikkita chapdagi va ikkita o‘ngdagi so‘zlarning lemmalari hamda ularning so‘z turkumlari quyidagicha bo‘ladi:

{shovqin,N;suruv,N;hurkitmoq,V;yubormoq }

Ya’ni, xususiyatlar vektori noaniq so‘zga nisbatan ikki oldingi va ikki keyingi so‘zlarning lemmalari hamda ularga mos so‘z turkumlaridan tashkil topadi.

Birga uchrash xususiyatlari maqsadli so‘z atrofida joylashgan so‘zlarga asoslanadi. Bu yondashuvda xususiyat sifatida so‘zlarning o‘zlari olinadi, ammo ularning so‘z turkumlari (POS) e’tiborga olinmaydi. Har bir xususiyatning qiymati ushbu so‘zlarning maqsad so‘zi kontekstida qanchalik tez-tez uchrashini bildiradi. Ushbu usulni samarali qo‘llash uchun odatda maqsadli so‘z yaqinida tez-tez uchraydigan, mazmun jihatdan muhim bo‘lgan kichik guruh so‘zlar xususiyat

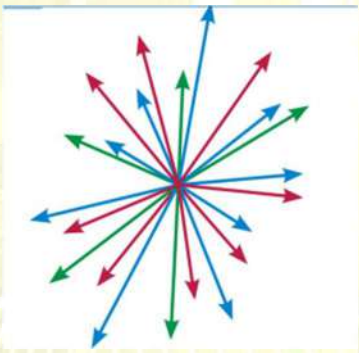
ISSN 3030-3907

sifatida tanlanadi. Masalan, “***ot***” so‘zi uchun uning turli ma’nolarini ifodalovchi jumlalardan eng ko‘p uchraydigan atrofdagi so‘zlar quyidagicha bo‘lishi mumkin:

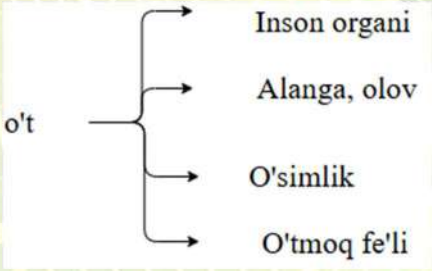
{tog‘, dara, chipor, toy, poyga, dala, ...}.

Masalan yangi kiritilgan *“Zotdor otlar orasida poyga uchun qabul davom etmoqda…”* gapidagi **ot** omonim so‘zning ma’nosini atrofidagi so‘zlar orqali aniqlash mumkin. Agar kontekst chegarasini 10 so‘z deb belgilansa, ushbu gapni mos keluvchi vektor shaklida tasvirlash mumkin.

Nazoratsiz o‘qitish algoritmlari. Bu yondashuvda ishlash jarayoni belgilangan teglarga ega bo‘lmagan ma’lumotlar to‘plamidan boshlanadi, ya’ni har bir ma’lumot birligi qaysi toifaga tegishli ekanligi noma’lum bo‘ladi. Faqat xususiyatlar mavjud bo‘lib, algoritm ushbu xususiyatlar asosida ma’lumotlardagi ichki teglarni mustaqil ravishda aniqlab, bir xil sinfga tegishli bo‘lgan elementlarni o‘zi guruhlashi kerak bo‘ladi.



3-rasm: Gapdagi omonim so‘zning kontekstdagi so‘zlari va ma’nolarining vektor ko‘rinishi



4-rasm: O‘t omonim so‘zining ma’nolari

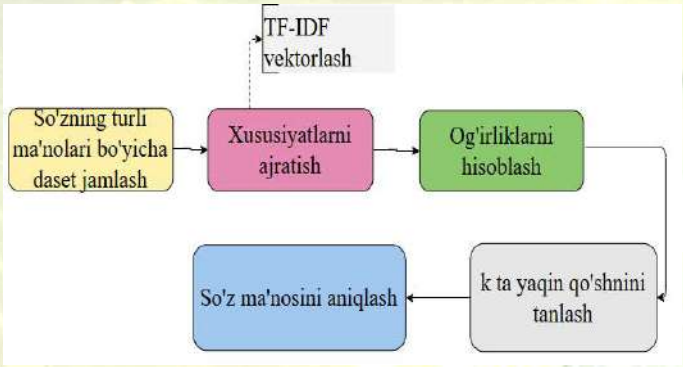
Mashina o‘rganish yondashuvlari hamda transformerlarga asoslangan modellar omonim so‘zlarning kontekstdagi ma’nosini aniqlashda katta imkoniyatlarni taqdim etadi. Ushbu tadqiqotda o‘zbek tilidagi jumlalar tarkibida omonim so‘zlarni semantik jihatdan farqlash uchun mashinaviy o‘rganish algoritmlaridan foydalanildi.

K-yaqin qo‘shni (K-NN) algoritmi. K-NN – Nazorat ostida o‘qitish (supervised) mashinaviy o‘rganish algoritmi bo‘lib, semantik o‘xshashlikka asoslangan tasniflash usulini qo‘llaydi [13; 16-23-b.]. Noaniq ma’noga ega bo‘lgan so‘z uchun k ta eng yaqin o‘xshash misollarni aniqlab, so‘zning kontekstdagi

ISSN 3030-3907

ma’nosini ana shu qo‘shnilar asosida belgilaydi. Ushbu jarayonning ketma-ketligi 5-rasmda ko‘rsatilgan.

Xususiyatlarni ajratib olish jarayoni ingliz tilida *Feature Extraction* nomi bilan yuritiladi [13; 16-23-b.]. So‘z ma’nosini aniqlash (WSD) tizimini ishlab chiqishda mos va dolzarb xususiyatlarni to‘g‘ri ajratib olish muhim ahamiyatga ega. Odatda xususiyatlar to‘plami bir nechta bosqichda amalga oshiriladi:



5-rasm: K-NN algoritmi yordamida so‘z ma’nosini aniqlash jarayoni

Term Frequency (TF) – bu usulda matnda eng ko‘p uchraydigan so‘zlar tanlab olinadi. TF asosida noaniq ma’noli so‘zning muayyan kontekstdagi turli ma’nolarini ifodalovchi jumlalarda eng tez-tez uchraydigan so‘zlar aniqlanadi va ular so‘zning mo‘ljallangan ma’nosini belgilashda yordam beradi [13; 16-23-b.].

K-yaqin qo‘shni (K-NN) algoritmidan foydalanilganda odatda quyidagi bosqichlar bajariladi:

1. **Ma’lumotlarni tayyorlash** – mavjud ma’lumotlar boshqa so‘zlar bilan ajratilgan birliklar (masalan, matndagi so‘zlar) ko‘rinishida bo‘ladi.
2. **Masofani o‘lchash.** So‘zlar o‘rtasidagi o‘xshashlik maxsus masofa metrikasi yordamida aniqlanadi. K-NN algoritmida bunday masofani hisoblashda odatda Evklid masofasi eng ko‘p qo‘llaniladi.
3. **k yaqin qo‘shnilarni topish.** Yangi so‘z uchun uning kontekstdagi eng yaqin k ta qo‘shni so‘z tanlab olinadi.
4. **O‘rganish va tasniflash.** Ana shu k ta qo‘shni asosida yangi so‘zning ma’nosi belgilanadi. Boshqacha qilib aytganda, yangi keltirilgan gapdagi noaniq so‘zning ma’nosi uning atrofidagi k qo‘shni so‘zlarning umumiy semantik mazmuni bilan bog‘lash orqali aniqlanadi [14; 415-439-b.].

So‘z ma’nosini aniqlash (WSD) jarayonida matn tarkibidagi so‘zlarning o‘rni va chastotasi muhim ahamiyat kasb etadi. Har bir so‘zning vazni quyidagi yondashuv asosida hisoblanishi mumkin:

Og‘irlikni belgilash strategiyasi. So‘zning vazni (og‘irligi) uning gapdagi joylashishi hamda kontekstda uchrash chastotasiga qarab aniqlanadi.

$$w(k, f_i) = \frac{N(k, f_i)}{N(k)} \qquad (1)$$

ISSN 3030-3907

bu yerda,

N(k, f_i)– so‘zning **k**-ma’nosi bilan bir kontekstdagi $\mathbf{f_i}$ xususiyat (feature) so‘zlari uchrashish chastotasi

N (k)– “**k**-ma’noli sinfga tegishli bo‘lgan barcha namunalarning umumiy soni”

1-Jadval

K-nn algoritmidan olingan natijalar tahlili

So‘z	**Kontextdagi so‘zlar bilan uchrashiga ko‘ra (%)**	**So‘zlarning chastotasi bo‘yicha aniqlash(%)**	**Ikkala xususiyat bo‘yicha (%)**
bo‘ydor	90.7	89.7	90.7
bo‘sh	76.8	75.8	76.8
bo‘yli	75.6	72.6	75.6
bog‘li	78.6	82.1	78.6
sirli	61.2	56.9	63.8
belli	70.1	65.7	71.1
O‘rtacha:	*75.5*	*73.8*	*76.1*

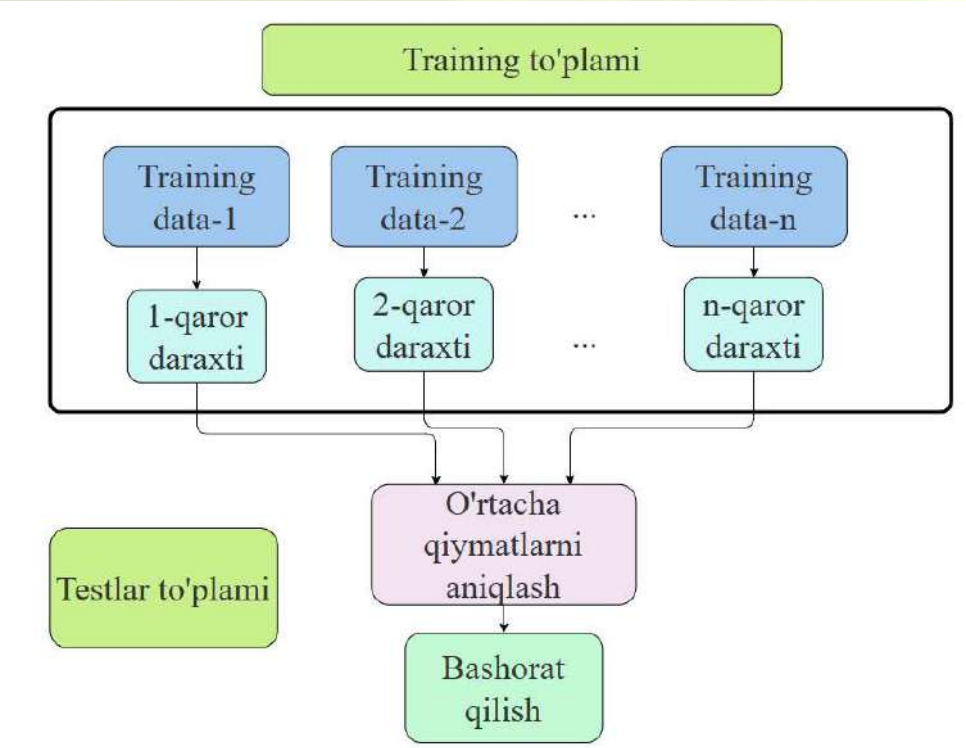
“Ushbu yondashuvning o‘rtacha aniqlilik ko‘rsatkichi **76,1 %** bo‘lib, u ilgari taklif qilingan usullarga nisbatan yaxshiroq samaradorlikni namoyon qildi.”

Tasodifiy o‘rmon (Random Forest). Ushbu algoritm bir nechta qaror daraxtlariga asoslanib ishlaydi va murakkab omonimlarni aniqlashda yaxshi natijalar beradi. Keling, Random Forest algoritmidan foydalanib matndagi omonim so‘zlarning ma’nosini belgilash jarayonini ko‘rib chiqamiz. Ushbu jarayon quyidagi asosiy bosqichlardan iborat:

- Omonim so‘zlar va ularning kontekstdagi jumlalari ma’lumotlar to‘plami (Excel fayli) orqali aniqlanadi.
- Model ushbu ma’lumotlar asosida o‘qitilib, so‘ngra yangi matnlardagi so‘z ma’nolarini aniqlashda qo‘llaniladi.
- Modelning aniqlik darajasi sinov orqali baholanadi.

Random Forest algoritmidan foydalangan holda model yaratish uchun avvalo omonim so‘zlarning har bir ma’nosiga oid jumlalar yig‘ib olinadi. Ushbu algoritm uchun mo‘ljallangan ma’lumotlar to‘plami 2-jadvalda keltirilgan shaklda tuzildi.

ISSN 3030-3907



6-rasm: Random Forest algoritmining ishlash bosqichlari

2-Jadval

Dataset uchun namuna

Omonim so‘z	**Kontekst**	**Ma’nosi**
Kasr	Kasr sonlarni o‘nli kasrga aylantirish bo‘yicha dars o‘tdik.	Siniq, parcha. Matematika termini; miqdor birligining qismi.
Kasr	U kasr tufayli yugurish musobaqasida qatnasha olmadi.	Shikastalik, nuqsonlilik
Kun	U kun nurlari ostida meditatsiya qildi.	Yerga issiqlik va nur taratib turuvchi planeta, quyosh, oftob.

Tahlil va natijalar. So‘z ma’nosini aniqlash dasturini ishlab chiqishda Python dasturlash tili imkoniyatlaridan foydalanildi. Modelni ishlab chiqish jarayonida pandas, scikit-learn, matplotlib. Seaborn kutubxonalaridan foydalanildi.

Modelni o‘qitish va baholash

– Excel faylidan jumlalar va ularning tegishli ma’nolari ajratib olindi.

– Ushbu jumlalar **CountVectorizer** yordamida vektor ko‘rinishiga o‘tkazildi.

– Ma’lumotlar to‘plami **train_test_split** funksiyasi yordamida 80/20 nisbatda o‘quv va sinov qismiga bo‘lindi.

– Modelni o‘qitish uchun **RandomForestClassifier()** algoritmi qo‘llanildi.

Modelni yaratish jarayonida jami 54 000 ta semantik jihatdan belgilangan jumlalar to‘plandi. Shulardan 33 200 tasi modelni o‘qitish uchun ishlatildi. Tayyor

ISSN 3030-3907

model 10 800 ta misolda sinovdan o‘tkazildi. Sinov bosqichidagi natijalarga ko‘ra, modelning samaradorligini baholash maqsadida **F1 ko‘rsatkichi** hisoblab chiqildi.

Sinov natijalari quyidagicha bo‘ldi:

- To‘g‘ri aniqlangan iboralar soni: 10 000 ta
- To‘g‘ri tasniflangan sinonimlar: 9 950 ta
- Noto‘g‘ri moslangan sinonimlar: 50 ta

Modelning umumiy ishlash sifatini baholashda ushbu ko‘rsatkichlar asosida **F1 metrikasi** qo‘llanildi.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Rec} \quad (2)$$

$$Precision = \frac{10000}{10000+50} = 0.99 \quad (3)$$

$$Recall = \frac{10000}{10000+995} = 0.50 \quad (4)$$

Hisoblangan **aniqlik (3)** va **to‘liqlik (4)** ko‘rsatkichlariga asoslanib, **(2)** qiymati 0,664 deb topildi. Ma’lumotlar 460 million so‘zdan iborat korpusdan olinib, u yangilik, ilmiy annotatsiyalar, og‘zaki nutq transkriptlari va badiiy matnlarni o‘z ichiga oladi. Hajmi jihatidan bu so‘z ma’nosini ajratib olish bo‘yicha ishlatilgan eng yirik o‘quv va sinov to‘plamlaridan biridir. Ushbu korpus zamonaviy transformerlarga asoslangan modeliga, xususan **Sense BERT** modeliga o‘qitildi va baholandi. Sense BERT asosida tayyorlangan model an’anaviy yondashuvlarga qaraganda ancha yuqori aniqlikni namoyon qildi. Shuningdek, *Random Forest* va *Sense BERT* algoritmlari uchun bir xil hajmdagi ma’lumotlar bazasi qo‘llanildi. *Sense BERT* modeli 92 % aniqlik ko‘rsatkichiga erishdi.

XULOSA. So‘z ma’nosini aniqlashning zamonaviy usullari bir nechta tajribada kuzatilgan lingvistik xususiyatlarga asoslanadi. Xususan, bitta kontekst yoki kollokatsiya ichida so‘z odatda yagona ma’noga ega bo‘lishi qonuniyatidan foydalaniladi. Algoritmlar ushbu xususiyatni samarali modellash orqali kollokatsion munosabatlar xilma-xilligini imkon qadar to‘liq ifodalashga intiladi. Mazkur tadqiqotda o‘zbek tilidagi omonim so‘zlarning kontekstdagi ma’nosini avtomatik aniqlash masalasi *Random Forest, K-Nearest Neighbors (KNN) va Sense BERT* algoritmlari yordamida ko‘rib chiqildi. Chunki omonim so‘zlar ishlatilgan kontekstga qarab turli ma’nolarni anglatishi mumkin, shu sababli ularning ma’nosini avtomatik aniqlash tabiiy tilni qayta ishlashdagi asosiy muammolardan biridir.

Tadqiqotning asosiy maqsadi foydalanuvchi kiritgan matndagi omonim so‘zning kontekstual ma’nosini aniqlashdan iborat bo‘ldi. Modelni yaratish jarayonida:

ISSN 3030-3907

• Omonim, jumla va ularning mos keluvchi ma’nosini o‘z ichiga olgan Excel formatidagi ma’lumotlar bazasi tuzildi.
• Matnlar tozalandi va lemmatizatsiya qilindi.
• Gaplar TF-IDF, Glove, Embedding vector kabi vektorizatsiya usullari yordamida raqamli vektorlarga aylantiriladi.
• Random Forest algoritmi quyidagi sabablarga ko‘ra modelni o‘qitish uchun tanlandi:
Klassifikatsiya vazifalarida yuqori aniqlikka erishadi.
Overfittingga chidamli.
Natijalarini talqin qilish oson.

Tayyorlangan yakuniy model foydalanuvchilar kiritgan matnlardagi omonim so‘zlarning kontekstdagi ma’nosini muvaffaqiyatli aniqlab berdi.

FOYDALANILGAN ADABIYOTLAR

1. Boltayevich, E.B., Mirdjonovna, H.S., and Ilxomovna, A.X. (2023). Methods for Creating a Morphological Analyzer. In: Zaynidinov, H., Singh, M., Tiwary, U.S., and Singh, D. (Eds.), *Intelligent Human Computer Interaction*. IHCI 2022, *Lecture Notes in Computer Science*, vol. 13741. Springer, Cham. https://doi.org/10.1007/978-3-031-27199-1_4
2. Oqila Abdullayeva. O‘zbek tili matnlarida sintaktik teg va teglash masalasi. O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi. 2024 Vol. 1 (6) B. 46-58.
3. Nikitina S. A. and Korzunina A. A. A neural network approach for determining the semantics of homonyms in russian-language texts. Vestnik komp iuternykh i informatsionnykh tekhnologii. February 2025, pp. 56-64
4. Ayazbayev, D.; Bogdanchikov, A.; Orynbekova, K.; Varlamis, I. Defining Semantically Close Words of Kazakh Language with Distributed System Apache Spark. *Big Data Cogn. Comput.* 2023, *7*, 160. https://doi.org/10.3390/bdcc7040160
5. El-Razzaz, M., Fakhr, M. W., & Maghraby, F. A. (2021). Arabic Gloss WSD Using BERT. *Applied Sciences*, *11*(6), 2567. https://doi.org/10.3390/app11062567
6. Zhong, Z., Ng, H.T.: It makes sense: a wide-coverage word sense disambiguation system for free text. In: ACL 2010—48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 78–83 (2010)
7. Zaripova D.A., Loukachevitch N.V. (2023) Approaches to Automatic Word Sense Disambiguation Based on Uneven Distribution of Word Senses in Corpus. Lomonosov Philology Journal. Series 9. Philology, no. 6, pp. 40–51
8. Daniil Alexeyevsky and Anastasiya V. Temchenko. 2016. Word Sense Disambiguation in Monolingual Dictionaries for Building Russian WordNet.

ISSN 3030-3907

In Proceedings of the 8th Global WordNet Conference (GWC), pages 9–14, Bucharest, Romania. Global Wordnet Association.

9. Elov, B., Abdurakhmonova, N., Axmedova, X., Xusainova, Z., Iskandarova, A., and Fattaxova, D. (2024). Designing Processes and Models of Semantical Differentiation for Polyfunctional Words in the Uzbek Contexts. In: Nguyen, N.T., et al. *Recent Challenges in Intelligent Information and Database Systems*. ACIIDS 2024, Communications in Computer and Information Science, vol. 2145. Springer, Singapore. https://doi.org/10.1007/978-981-97-5934-7_26

10. Boltayevich, E.B., Ilxomovna, A.X., Hakim Qizi, P.M., and Uktambay O‘g‘li, K.N. (2023). Semantic Differentiation of Uzbek Homonyms Using the Lesk Algorithm. *2023 8th International Conference on Computer Science and Engineering (UBMK)*, Burdur, Turkiye, pp. 137–140. https://doi.org/10.1109/UBMK59864.2023.10286666

11. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

12. Mihalcea, R., and Moldovan, D. (2002). Word sense disambiguation with pattern learning and automatic feature selection. *Natural Language Engineering, 8*(4), 349–370. https://doi.org/10.1017/S1351324902002991

13. Abdul-Rauf and Holger Schwenk. On the use of Comparable Corpora to improve SMT performance Sadaf. 12th Conference of the European Chapter of the ACL, pages 16–23, Athens, Greece, 30 March – 3 April 2009.pp 16-23

14. Gale, K. Church, and D. Yarowsky.: A Method for Disambiguating Word Senses in a Large Corpus. Computers and Humanities, vol. 26, pp. 415-439 (1992).

ISSN 3030-3907