MATN DAVRINI AVTOMATIK ANIQLASH USULLARI

Alaev Ruhillo Habibovich,
t.f.f.d. (PhD), dotsent
alayev_r@nuu.uz
O‘zMU

Kenjayeva Marjona,
I bosqich magistrant
mkenjayeva274@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu tezis matni matnlarning yozilgan davrini avtomatik aniqlash masalasida qo‘llaniladigan zamonaviy yondashuvlarni tahlil qiladi. Hujjatlarda aniq sanalar ko‘rsatilmagan hollarda, ularning davrini tilning diaxronik o‘zgarishlariga asoslanib topish mumkin. Tadqiqotda n-grammalarga asoslangan an’anaviy stilometriya, chuqur o‘rganish va katta til modellari samaradorligi qiyoslanadi. Shuningdek, o‘zbek tilining tarixiy bosqichlari (Qadimgi turkiy, Chig‘atoy, Zamonaviy o‘zbek tili) va milliy korpuslar yordamida mahalliy matnlar davrini tasniflash istiqbollari yoritilgan.

Kalit so‘zlar: *matn davrini aniqlash, diaxronik lingvistika, kompyuter lingvistikasi, Transformer, TALM, LLM, o‘zbek tili korpusi, stilometriya, ekstralingvistik teglash.*

Abstract. This thesis text analyzes modern approaches applied in the task of Automatic Text Dating. In cases where explicit dates are not provided in documents, their time period can be determined based on diachronic language changes. The research compares the effectiveness of traditional n-gram-based stylometry, deep learning, and large language models. Additionally, it highlights the prospects of classifying the age of local texts using the historical stages of the Uzbek language (Old Turkic, Chagatai, Modern Uzbek) and national corpora .

Keywords: *automatic Text Dating, diachronic linguistics, computational linguistics, Transformer, TALM, LLM, Uzbek language corpus, stylometry, extralinguistic tagging.*

Kirish. So‘nggi yillarda axborot texnologiyalarining jadal rivojlanishi natijasida turli sohalarga oid katta hajmdagi matnli ma’lumotlar shakllanmoqda. Elektron kutubxonalar, raqamli arxivlar, ilmiy bazalar hamda internet tarmoqlarida saqlanayotgan matnlarni tartiblash, izlash va tahlil qilish dolzarb masalaga aylangan. Ushbu jarayonda matnlarning yozilgan vaqtini aniqlash, ya’ni matn davrini belgilash muhim ilmiy-amaliy ahamiyat kasb etadi. Matn davrini avtomatik aniqlash matnning qaysi tarixiy davrga yoki vaqt oralig‘iga tegishli ekanligini lingvistik, statistik va hisoblash usullari yordamida aniqlash jarayonidir. Ushbu yo‘nalish tabiiy tilni qayta ishlash (Natural Language Processing, NLP), sun’iy intellekt hamda hisoblash lingvistikasining muhim tadqiqot sohalaridan biri hisoblanadi.

Til doimiy ravishda rivojlanib boruvchi tizim bo‘lib, vaqt o‘tishi bilan leksik birliklar, grammatik qurilmalar, uslubiy xususiyatlar va semantik tuzilmalar o‘zgaradi. Shu sababli ma’lum bir davrga oid matnlar o‘ziga xos lingvistik belgilar bilan ajralib turadi. Masalan, tarixiy matnlarda arxaik so‘zlar va eski grammatik shakllar uchrasa, zamonaviy matnlarda yangi terminologiya va innovatsion ifodalar ko‘proq qo‘llaniladi. Matn davrini aniqlash tarixiy manbalarni raqamlashtirish, mualliflikni aniqlash, plagiatni tekshirish, elektron arxivlarni tashkil etish hamda axborot qidiruv tizimlarining samaradorligini oshirishda muhim rol o‘ynaydi. Ayniqsa, o‘zbek tilidagi tarixiy va zamonaviy matnlarni avtomatik tasniflash bo‘yicha tadqiqotlar yetarli darajada rivojlanmaganligi ushbu mavzuning dolzarbligini yanada oshiradi.

Asosiy qism

Matn davrini avtomatik aniqlash jarayoni matnlarni kompyuter yordamida tahlil qilish hamda ularning yozilgan vaqtiga xos lingvistik belgilarni aniqlashga asoslanadi. Ushbu jarayonda matn dastlab qayta ishlanadi, ya’ni ortiqcha belgilar tozalanadi, so‘zlar ajratiladi va matn tahlil uchun qulay ko‘rinishga keltiriladi. Matnni dastlabki qayta ishlash bosqichi keyingi hisoblash jarayonlarining aniqligini

ta’minlashda muhim ahamiyat kasb etadi [5:37-41]. Tilning tarixiy rivojlanishi natijasida har bir davrga xos leksik, grammatik va uslubiy xususiyatlar shakllanadi. Vaqt o‘tishi bilan ayrim so‘zlar iste’moldan chiqadi, yangi terminlar paydo bo‘ladi, gap tuzilishi hamda ifoda uslubi o‘zgaradi. Shu sababli matn tarkibidagi so‘zlar chastotasi, arxaik birliklarning mavjudligi, morfologik shakllar va sintaktik konstruktsiyalar matn davrini aniqlashda asosiy ko‘rsatkich sifatida xizmat qiladi. Lingvistik belgilarni aniqlash orqali matnning qaysi tarixiy davrga mansubligi haqida muayyan xulosaga kelish mumkin bo‘ladi.

Matn davrini aniqlash jarayoni, avvalo, matnni dastlabki qayta ishlash bosqichidan boshlanadi. Ushbu bosqichda matn kompyuter tahliliga mos shaklga keltiriladi: ortiqcha belgilar olib tashlanadi, so‘zlar tokenlarga ajratiladi, funksional birliklar filtrlanadi hamda morfologik normalizatsiya amalga oshiriladi. Matnni standartlashtirish keyingi statistik va algoritmik jarayonlarning samaradorligini oshiradi [7:32-35]. Til tizimi vaqt o‘tishi bilan o‘zgarib boruvchi dinamik hodisa bo‘lib, har bir tarixiy davr o‘ziga xos lingvistik belgilar bilan ajralib turadi. Leksik qatlamdagi o‘zgarishlar, yangi terminlarning paydo bo‘lishi, ayrim so‘zlarning eskirishi, grammatik shakllarning evolyutsiyasi hamda stilistik normalarning almashuvi matn davrini aniqlashda asosiy mezon sifatida xizmat qiladi. Masalan, tarixiy matnlarda arxaik birliklar, diniy va klassik adabiyotga xos iboralar uchrasa, zamonaviy matnlarda texnologik, ijtimoiy va global jarayonlarni ifodalovchi terminlar ustunlik qiladi [2:63]. Matn davrini avtomatik aniqlashda statistik tahlil usullari muhim o‘rin tutadi. Statistik yondashuvlar matndagi so‘zlar chastotasi, n-gramm modellari hamda ehtimollik taqsimotlari asosida ishlaydi. Ushbu metodlar yordamida turli davrlarga tegishli matnlar o‘rtasidagi lingvistik farqlar aniqlanadi. Klassifikatsiya jarayonida mashinaviy o‘qitish algoritmlaridan foydalanish matnni avtomatik ravishda ma’lum vaqt oralig‘iga ajratish imkonini beradi [1:72-74].

1-rasm. Matn davrini avtomatik aniqlash jarayoni

Zamonaviy tadqiqotlarda Naive Bayes, Support Vector Machine, Decision Tree hamda Random Forest algoritmlari matn klassifikatsiyasida samarali natijalar ko‘rsatmoqda. Ushbu algoritmlar oldindan belgilangan matnlar korpusi asosida o‘qitilib, yangi matnning davrini aniqlash imkonini yaratadi. Mashinaviy o‘qitish yondashuvlari inson omiliga bog‘liqlikni kamaytirib, katta hajmdagi ma’lumotlarni tezkor tahlil qilishni ta’minlaydi [4:65]. So‘nggi yillarda chuqur o‘rganish (Deep Learning) texnologiyalarining rivojlanishi matn davrini aniqlash jarayonida yangi bosqichni boshlab berdi. Neyron tarmoqlar, xususan RNN va LSTM modellar hamda transformer arxitekturasiga asoslangan til modellari matn kontekstini chuqur o‘rganish imkoniyatiga ega. Ushbu modellar matndagi yashirin semantik aloqalarni aniqlash orqali davrni yuqori aniqlik bilan bashorat qiladi [6:96]. Matn davrini avtomatik aniqlash natijalari raqamli gumanitar tadqiqotlar, elektron kutubxonalar yaratish, tarixiy qo‘lyozmalarni tasniflash, mualliflik ekspertizasi, plagiatni aniqlash hamda aqlli axborot qidiruv tizimlarini ishlab chiqishda keng qo‘llanilmoqda. Shu sababli mazkur yo‘nalish sun’iy intellekt va hisoblash lingvistikasining istiqbolli ilmiy yo‘nalishlaridan biri sifatida qaraladi [3:83-87].

Raqamli arxivlar, kutubxonalar va internet resurslari misli ko‘rilmagan darajada kengayib borayotgan bir paytda, sanasi noma’lum tarixiy hujjatlarning yozilgan davrini avtomatik aniqlash muammosi kompyuter lingvistikasining dolzarb vazifasiga aylandi. Bu jarayon matnning leksik, morfologik va sintaktik qurilishida yuz bergan diaxronik o‘zgarishlarni tahlil qilish orqali amalga oshiriladi. Jonli til

ijtimoiy omillar ta’sirida doimiy ravishda o‘zgarishda bo‘lgani uchun neologizmlar, arxaizmlar, imlo va grammatik siljishlar kompyuter modellari uchun hujjat davrini belgilovchi asosiy “markerlar” bo‘lib xizmat qiladi.

Stilometrik va an’anaviy mashinali o‘qitish usullari. Dastlabki yondashuvlar asar muallifining individual uslubini yoki davr qoidalarini matematik tahlil qiluvchi stilometriyaga tayangan. Belgilar, so‘zlar yoki grammatik qoidalar ketma-ketligini o‘rganish matn davrini topishda eng kuchli an’anaviy vosita hisoblanadi. Xususan, belgilar n-grammasi asrlar davomida o‘zgargan qadimiy imlo qoidalarini tutishda juda samarali. So‘zlarning vaqt o‘qi bo‘ylab matnlarda takrorlanish chastotasi o‘rganiladi. Bu usul qisqa va uzun muddatli tebranishlarni (masalan, haftaning ma’lum kunlari yoki yillar bilan bog‘liq so‘zlarni) ajratib olish imkonini beradi. N-grammalar so‘zlarning chuqur semantik kontekstini tushuna olmagani sababli, NLP sohasida neyron tarmoqlari ommalashdi. Konvolyutsion (CNN) tarmoqlar matndagi mahalliy uslubiy naqshlarni fosh qilsa, Rekurrent (RNN/LSTM) tarmoqlar jumlalar orasidagi uzun masofali mantiqiy bog‘lanishlarni xotirasida saqlab, matnning tarixiy fonini aniqroq baholaydi. Biroq, standart BERT va RoBERTa modellari diaxronik semantik siljishlarni (bitta so‘zning turli asrlarda turlicha ma’no anglatishini) hisobga olmasdan o‘qitilgani bois, tarixiy matnlarda biroz oqsaydi.

Xulosa. Matnlarning yozilgan davrini avtomatik aniqlash texnologiyalari an’anaviy stilometriyadan boshlanib, bugungi kunda TALM va katta til modellari kabi kognitiv darajadagi intellektual yechimlargacha bo‘lgan yo‘lni bosib o‘tdi. O‘zbek tili uchun bu jarayonlar endigina jadal pallasiga kirmoqda. Milliy korpuslarni ekstralingvistik teglash hamda mashinali tahlil vositalarini chuqur modellar bilan integratsiya qilish orqali kelajakda O‘zbekistonning yirik raqamli arxivlarini xronologik jihatdan avtomatik tizimlashtiruvchi mukammal va aqlli dasturlar yaratilishi muqarrardir. Matn davrini avtomatik aniqlash zamonaviy tabiiy

tilni qayta ishlash va hisoblash lingvistikasi sohalarining dolzarb yo‘nalishlaridan biri hisoblanadi. Olib borilgan tahlillar shuni ko‘rsatadiki, matn tarkibidagi lingvistik, statistik va semantik belgilar uning qaysi tarixiy davrga tegishli ekanligini aniqlashda muhim rol o‘ynaydi. Tilning vaqt davomida o‘zgarishi natijasida leksik birliklar, grammatik tuzilmalar va uslubiy xususiyatlar transformatsiyaga uchraydi, bu esa matnlarni avtomatik tasniflash imkonini beradi.

Tadqiqot jarayonida an’anaviy statistik usullar bilan bir qatorda mashinaviy o‘qitish va chuqur o‘rganish yondashuvlari ham samarali ekanligi aniqlandi. Ayniqsa, neyron tarmoqlar va transformer modellari matn kontekstini chuqur tahlil qilish orqali yuqori aniqlikdagi natijalarni ta’minlamoqda. Shu bilan birga, ushbu usullar katta hajmdagi ma’lumotlar va hisoblash resurslarini talab qilishi bilan ajralib turadi. Umuman olganda, matn davrini avtomatik aniqlash tizimlarini rivojlantirish tarixiy matnlarni tahlil qilish, raqamli arxivlarni yaratish, mualliflikni aniqlash va axborot qidiruv tizimlarini takomillashtirishda katta amaliy ahamiyatga ega. Kelajakda ushbu yo‘nalishda o‘zbek tiliga moslashtirilgan korpuslar va yanada aniqroq modellar yaratish dolzarb vazifa bo‘lib qoladi.

Foydalanilgan adabiyotlar ro‘yxati

1. Aggarwal C.C. Machine Learning for Text. – Cham: Springer, 2018. – 300 p.
2. Bird S., Klein E., Loper E. Natural Language Processing with Python. – Sebastopol: O’Reilly Media, 2020. – 504 p.
3. Boltayev B. E., Xolmo‘minova A. O. O‘zbek tilini avtomatik qayta ishlash va lingvistik tahlil masalalari. – Toshkent: Toshkent davlat universiteti nashriyoti, 2022. – 250 b.
4. Goldberg Y. Neural Network Methods for Natural Language Processing. – San Rafael: Morgan & Claypool, 2017. – 309 p.

5. Jurafsky D., Martin J. H. Speech and Language Processing. – Hoboken: Pearson, 2023. – 600 p.

6. Jo‘rayev A. A. Hisoblash lingvistikasi asoslari. – Toshkent, 2021. – 200 b.

7. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. – Cambridge: MIT Press, 2021. – 680 p.