III SHO‘BA. TABIIY TILNI QAYTA ISHLASH (NLP)

UDK: 004.912

MATNNI TAHLIL QILISH BO‘YICHA ILG‘OR TAJRIBALAR

Mamatov Narzillo Solidjonovich,
Texnika fanlari doktori, professor
m_narzullo@mail.ru
“Toshkent irrigasiya va qishloq xo‘jaligini mexanizasiyalash muhandislari”Milliy tadqiqot universiteti

Nuritdinov Nurbek Davlataliyevich,
nuritdinovnurbek85@gmail.com
Namangan muhandislik-qurilish instituti, tayanch-doktorant

Annotatsiya: Ushbu maqolada matnlarni tahlil qilishning turli fan sohalarida qo‘llaniladigan asosiy usullari tahlil qilinadi va ularning kompyuterlashtirish darajasi bo‘yicha tasnifi keltiriladi. Tadqiqotda tematik tahlil, kontent tahlili, lug‘at yondashuvlari, so‘zlar sumkasi (BOW) modeli, boshqariladigan va boshqarilmaydigan o‘rganish usullari, shuningdek tabiiy tilni qayta ishlash (NLP) texnikalari qamrab olingan. Har bir yondashuvning nazariy asoslari, amaliy qo‘llanilishi, afzalliklari va cheklovlari muhokama qilinadi. Maqolaning maqsadi – matn tahlilining turli metodologik yondashuvlarini tizimlashtirish va tadqiqotchilar uchun ularning imkoniyatlari hamda chegaralarini aniqlab berishdan iborat. Tadqiqot natijalari matn tahlil usullarini tanlashda asos bo‘lishi, ilmiy, ijtimoiy va texnologik sohalarda matnga asoslangan tadqiqotlar samaradorligini oshirishga xizmat qilishi mumkin.

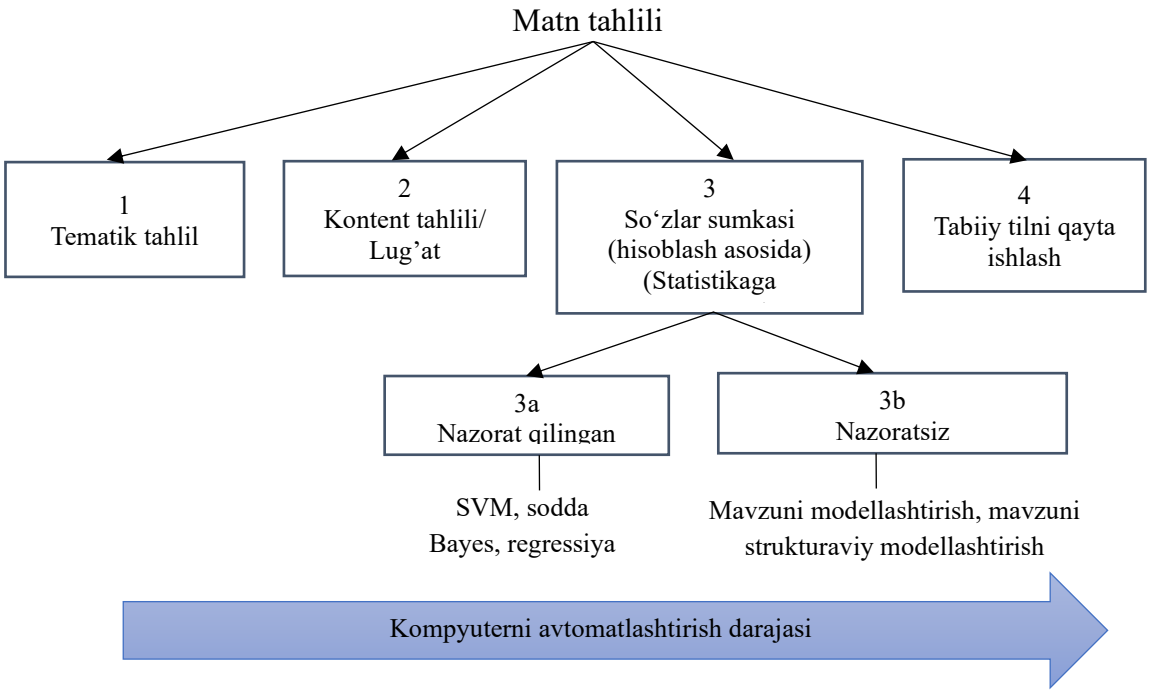
Аннотация: В данной статье рассматриваются основные методы анализа текстов, применяемые в различных научных дисциплинах, и приводится их классификация в зависимости от степени компьютеризации. В исследование включены тематический анализ, контент-анализ, лексические подходы, модель «мешка слов» (BoW), методы контролируемого и неконтролируемого обучения, а также технологии обработки естественного языка (NLP). Обсуждаются теоретические основы, практическое применение, преимущества и ограничения каждого подхода. Цель статьи — систематизировать различные методологические подходы к анализу текста и определить их возможности и ограничения для исследователей. Результаты исследования могут послужить основой для выбора методов анализа текста, а также способствовать повышению эффективности текстоориентированных исследований в научной, социальной и технологической сферах.

Abstract: This article examines the main methods of text analysis used across various academic disciplines and provides their classification based on the degree of

computerization. The study includes thematic analysis, content analysis, lexical approaches, the bag-of-words (BoW) model, supervised and unsupervised learning methods, as well as natural language processing (NLP) technologies. Theoretical foundations, practical applications, advantages, and limitations of each approach are discussed. The aim of the article is to systematize different methodological approaches to text analysis and to identify their capabilities and limitations for researchers. The research findings can serve as a basis for selecting appropriate text analysis methods and contribute to enhancing the effectiveness of text-based research in scientific, social, and technological domains.

Kalit so‘zlar: *Matn tahlili, Tematik tahlil, Kontent tahlili, So‘zlar sumkasi modeli, Tabiiy tilni qayta ishlash, Lug‘atga asoslangan yondashuv, Fanlararo tahlil, Semantik tahlil, Statistik yondashuvlar, Sun’iy intellekt.*

Ushbu maqolada matnlarni tahlil qilishning turli fan sohalarida qo‘llaniladigan turli xil matn tahlil usullarini tizimlashtiradi va muhokama qiladi. Matnni tahlil qilish ko‘p shakllarni oladi, ularning har biri o‘z taxminlari, afzalliklari va cheklovlariga ega[1]. Muayyan fan doirasida qo‘llaniladigan o‘ziga xos texnikalar asosan noma’lum yoki boshqa fanlar tomonidan qo‘llanilmaydi, bu esa har bir texnikaning o‘ziga xos xususiyatlaridan xabardor bo‘lmasligiga olib kelishi mumkin. 1-rasmda kompyuterni avtomatlashtirish darajasi bo‘yicha guruhlangan turli fanlarda qo‘llaniladigan eng mashhur matn tahlil usullarining tasnifi keltirilgan. Ushbu turkumlash tadqiqotchilarga matn ma’lumotlarini tahlil qilish uchun mavjud vositalarni ko‘rish imkonini beruvchi tashkiliy tuzilmani taqdim etadi [2].



1-rasm. Matnni tahlil qilishning umumiy usullarining tasnifi tematik tahlil.

1-rasmning 1-bandiga nazar tashlaydigan bo‘lsak, tematik tahlil odatda asosli nazariya metodologiyalari bilan bog‘liq bo‘lgan matn tahliliga sharhlovchi yondashuvdir. Asoslangan nazariya va tematik tahlilning maqsadi odamlarning o‘z dunyolariga xos bo‘lgan ma’no va ma’nolarga asoslangan nazariyani yaratish va takomillashtirishdir. Binobarin, ma’lumotlar ko‘pincha intervyu stenogrammasi, ishtirokchilar kuzatuvlari qaydlari va arxiv matni, jumladan hujjatlar, veb-saytlar va elektron pochta xabarlari shaklida taqdim etiladi. Tematik tahlil ko‘pincha tashkiliy fanlar va aloqa fanlarida qo‘llaniladi [3].

Tematik tahlil takrorlanuvchi jarayonni o‘z ichiga oladi, bunda tadqiqotchi matndan kelib chiqadigan bir qator kodlar va toifalarni ishlab chiqadi. Umuman olganda, toifalar tahlil boshlanishidan oldin noma’lum, nazariyani takomillashtirish zarur bo‘lgan hollar bundan mustasno; bunday hollarda ma’lumotlarni tahlil qilish adabiyot va ma’lumotlarni doimiy ravishda taqqoslashni o‘z ichiga oladi. Tadqiqotchi ishtirokchilarning o‘z tilidan ("birinchi tartib kodlari" yoki "ochiq kodlash" deb ataladi) boshlanadi va keyin o‘xshash kodlarni toifalarga ("ikkinchi tartib kodlari" yoki "eksenli kodlash" deb ataladi). NVivo va ATLAS kabi kompyuter dasturlari bunday kodlar va toifalarni tashkil qilishni osonlashtirishga yordam beradi, garchi matn toifalari odatda inson kodlashidan olingan toifalarning operasion ta’riflariga bog‘liq bo‘lsada, shuning uchun ko‘pincha kompyuterni avtomatlashtirish darajasi juda kam yoki umuman yo‘q [4].

Garchi tematik tahlil tadqiqotchilarga unchalik ma’lum bo‘lmagan hodisalar va tushunchalarni tushunish imkonini beradi, ammo bu usul bilan bog‘liq bir qator xarajatlar mavjud. Kodlash yarim yoki to‘liq avtomatlashtirilmaganligi va toifalar noma’lum bo‘lganligi sababli matnni o‘qish, turkumlashtirish va talqin qilish ko‘p vaqt talab etadi, bu esa katta hajmdagi ma’lumotlarni (masalan, 100 000 varaq transkript) tahlil qilishni deyarli imkonsiz qiladi. Bundan tashqari, tematik tahlil xatolar va tadqiqotchilarning noto‘g‘riligiga nisbatan zaif bo‘lishi mumkin. Bundan tashqari, tematik tahlil odatda boshqa yondashuvlarga qaraganda qiziqish sohasi haqida ko‘proq bilim talab qiladi.

Kontentni tahlil qilish lug‘atga asoslangan usullar. Kontentni tahlil qilish va boshqa lug‘at usullari (1-rasm, 2-band) ko‘pincha ma’lum bir matndagi so‘zlar yoki iboralar chastotasini hisoblash orqali amalga oshiriladi. SHuning uchun, lug‘at usullari yordamida sifatli ma’lumotlar miqdoriy yo‘naltirilgan tadqiqot savollariga javob berish uchun ishlatilishi mumkin, chunki matn ko‘pincha so‘z chastotalarini hisoblash uchun qisqartiriladi. Oldindan talab qilinadigan inson bilimlari miqdori, shuningdek, kompyuterni avtomatlashtirish, so‘zlar ro‘yxati ma’lumotlardan kelib chiqadimi yoki so‘zlar ro‘yxati nazariya yordamida apriori aniqlanganiga qarab o‘zgaradi. Tematik tahlilga o‘xshab, kompyuter dasturlari kontentni tahlil qilish jarayonida yordam berishi mumkin. DICTION kabi dasturlar turkumlash lug‘atidan foydalangan holda matnni avtomatik ravishda baholaydi (ya’ni, operativ ta’riflar o‘rniga so‘zlar yoki n-grammalarga asoslangan mavzularni aniqlash). NVivo yoki ATLAS kabi boshqa dasturlar tematik tahlilga o‘xshab qo‘llanilishi mumkin, bunda kodlash va toifalash ma’lumotlarni tashkil etish dasturi yordamida qo‘lda amalga oshiriladi [5].

So‘zlar sumkasi va tabiiy tilni qayta ishlash.

Oxirgi ikkita usul - matnni tahlil qilishda so‘zlar sumkasi (BOW; ya’ni sanashga asoslangan; 1-rasm, 3-blok) va tabiiy tilni qayta ishlash (NLP; 1-rasm, 4-blok) yondashuvlari. Ushbu matnni tahlil qilish vositalari kompyuter dasturlari turli matnlarni tasniflash va toifalarga ajratish vazifalarini o‘rganadigan va ularni yarim yoki to‘liq avtomatik ravishda bajarishi mumkin bo‘lgan mashinani o‘rganishni o‘z ichiga olishi mumkin. Sun’iy intellekt inson tajribasi darajasiga (yoki undan ham yaxshiroq) erishishi mumkin. SHu bilan birga, inson tajribasini sun’iy intellekt bilan birlashtirish sinergetik ta’sir ko‘rsatishi mumkin.

BOW yondashuvlari matnni soddalashtirish va qisqartirish uchun asosan informatika fanlarida qo‘llaniladigan texnikalar guruhidir[6]. Bu usullar hujjatdagi so‘zlarning tartibi muhim emasligini nazarda tutadi; ya’ni hujjatlar "so‘zlar sumkasi" sifatida qaraladi. So‘z tartibini e’tiborsiz qoldirish hujjatda so‘zlarning necha marta paydo bo‘lishini hisoblash orqali ma’lumotlarga statistik xususiyatlarni qo‘shadi. Olingan ma’lumotlar strukturasi hujjat atamasi matrisasi deb ataladi, unda

fayl ma’lumotlar matrisasining har bir satri hujjatni, ustunlar esa hujjatda qo‘llanilgan alohida atamalarni ifodalaydi. BOW usullari so‘z tartibini e’tiborsiz qoldirganligi sababli, ular statistik xususiyatlarni (ya’ni, o‘rnini bosish) ta’minlaydi. Natijada, bu usullar hujjatlarni toifalarga bo‘lish kabi makro vazifalarni bajarishda yaxshi va katta hajmdagi ma’lumotlarni qayta ishlashga qodir. Misol uchun, spam-filtrlar ba’zi elektron pochta xabarlarini mazmuniga ko‘ra spam sifatida tasniflash uchun Baes ehtimollik modellariga asoslangan BOW yondashuvidan foydalanadi. Biroq, BOW usullari matnning semantik ma’nosini aniqlash kabi mikro darajadagi vazifalar uchun mos emas.

BOW yondashuvlarining ikki turi mavjud: nazorat ostida (1-rasm, 3a blok) va nazoratsiz
(1-rasm, 3b blok). Ushbu usullar va ularni qo‘llash tadqiqot kontekstiga qarab farq qiladi. Boshqariladigan usullarda tadqiqotchi nimani qidirayotganini oldindan biladi. Aniqrog‘i, tadqiqotchi dasturga kirish ma’lumotlarini (bu holda matn) ham, chiqish ma’lumotlarini ham (masalan, matn muallifini aniqlash) beradi va tizim ular orasidagi munosabatni ko‘rsatish uchun algoritm yaratadi. Tadqiqotchilar Mosteller va Uolles 12 ta munozarali federalistik hujjatlarning (Jeyms Madison yoki Aleksandr Hamilton) muallifligini bashorat qilish uchun oddiy Baescha so‘z ehtimollaridan foydalangan holda ushbu yondashuvning eng dastlabki misollaridan birini taqdim etdilar. Bugungi kunda Naive Baes va Support Vector Machines (SVMs) kabi usullar matn tahlili uchun ishlatiladigan mashhur nazorat ostidagi algoritmlardir [7].

Shu bilan bir qatorda, nazoratsiz algoritmlar tematik tahlilga o‘xshash so‘z klasterlari va ma’lumotlardan kelib chiqadigan mavzularni aniqlaydi. Biroq, mavzu tahlilidan farqli o‘laroq, mavzuni modellashtirish muhim mavzularni aniqlash uchun yuqori darajada (to‘liq bo‘lmasa ham) avtomatlashtirilgan yondashuvdan foydalanadi. Tahlil qilish uchun kamroq vaqt va matn haqida kamroq ma’lumot talab etiladi. Shu sababli, mavzuni modellashtirish katta hajmdagi ma’lumotlarni tahlil qilish uchun javob beradi, garchi inson tushunchasi paydo bo‘lgan mavzularni sharhlash uchun hali ham muhim. Mavzuni modellashtirish mavzuni tahlil qilish (ya’ni, inson tushunchasi) va mashinani o‘rganish (ya’ni, katta hajmdagi matnni tezda tahlil qilish) afzalliklaridan foydalanadi.

Nihoyat, NLP odatda matn tahlilining eng avtomatlashtirilgan shaklidir. Bu usul odamlarning tilni qanday tushunishi va qayta ishlashini modellashtiradi. Masalan, NLP usullari jumladagi so‘zlarning nutq qismlarini (masalan, otlar, sifatlar va boshqalar) belgilash imkonini beradi. Hujjatlarni bir tildan boshqa tilga tarjima qiling va hatto so‘zlarning ma’nosini aniqlashtirish uchun jumla kontekstidan foydalaning. Shuning uchun, BOW yondashuvidan farqli o‘laroq, NLP so‘z tartibini muhim deb hisoblaydi. CHuqur o‘rganish va multimodallik (ya’ni, matn va tasvirlarni birlashtirish) kabi ilg‘or usullardan foydalangan holda his-tuyg‘ularni

tahlil qilish o‘quv to‘plamlaridan foydalanishda NLPning mashhur shakllaridan biridir. Ushbu maxsus tahlil matnga nisbatan umumiy munosabat, hissiyot yoki fikrni ijobiy, salbiy yoki neytral deb tasniflaydi. Mavzuni tahlil qilishdan farqli o‘laroq, NLP to‘liq avtomatlashtirilgan jarayon bo‘lib, shuning uchun inson tushunishi va talqinini kam yoki umuman talab qilmaydi. Bundan tashqari, insonni kodlashni talab qiladigan usullar (masalan, tematik tahlil) bilan solishtirganda, NLP boshqa yondashuvlarga qaraganda ancha tez va tizimli. Masalan, kompyuter fanlari, informatika, tilshunoslik va psixologiya sohasidagi tadqiqotchilar NLP dan matnni qazib olish vositasi sifatida foydalanadilar [8].

Tekshirish, baholash va izohlash uchun talab qilinadigan mavzuni modellashtirish bosqichlari [9]

1-bosqich: Gipoteza va savollarni shakllantirish

(a) Gipotezalar tadqiqot muammolari. Mavzuni modellashtirishda gipotezalar va tadqiqot muammolari boshqa har qanday tadqiqotga o‘xshash ishlab chiqiladi.

2-qadam: Dizayn va ma’lumotlarni yig‘ish

(a) ma’lumotlar turi. Ma’lumotlar turlariga misollar orasida ochiq so‘rov javoblari va arxiv ma’lumotlari, masalan, veb-sayt matni, bosh direktorlarning xatlari, hamkasblar o‘rtasidagi elektron pochta xabarlari va ijtimoiy media (masalan, Twitter, Facebook) kiradi.

(b) Namuna hajmi. Namuna hajmi farazlar va tadqiqot savollarining nazariy va uslubiy mulohazalari bilan bog‘liq. Masalan, nazariya ma’lum bir qiziqish populyasiyasini ko‘rsatishi mumkin yoki korrelyasion tahlil o‘tkazish istagi namuna hajmini aniqlash uchun quvvat tahlilini talab qilishi mumkin. Tadqiqotchilar hujjat darajasidagi tahlilni ham ko‘rib chiqishlari kerak.

(c) yozish sifati. Yaxshiroq yozish odatda tahlilda yordam beradi. Biroq, hatto sifatsiz matnni ham tahlil qilish mumkin (masalan, Twitter ma’lumotlari).

(d) javoblar davomiyligi. Ko‘proq matn odatda kamroqdan yaxshiroqdir. Minimal so‘zlar soni bo‘lmasada, har bir hujjat uchun kamida 200 belgi tavsiya etiladi.

(e) kovariasiyalar. Turli mavzular paydo bo‘ladimi yoki quyi toifaga qarab har xil muhokama qilinadimi yoki yo‘qligini tushunish uchun "kovariatlar" sifatida sinovdan o‘tkazilishi mumkin bo‘lgan nazariy va uslubiy moderatorlarni hisobga olish kerak.

3-qadam: Oldindan davolash

(a) noto‘g‘ri yozuvlar. Minimal so‘zlar soniga to‘g‘ri kelmaydigan yoki tegishli matnni o‘z ichiga olmagan javoblar kabi noto‘g‘ri yozuvlarni o‘chirish haqida o‘ylashingiz mumkin.

(b) tokenizasiya. Tokenizasiya bosqichi jumlalarni alohida so‘zlarga yoki "tokenlarga" qisqartirishni o‘z ichiga oladi.

(c) tozalash. Tozalash bosqichi kichik harf belgilarini yaratish, bo‘shliqlar va tinish belgilarini olib tashlashni o‘z ichiga oladi.

(d) so‘zlarni to‘xtating. To‘xtash so‘zlarini olib tashlash tavsiya etiladi. To‘xtash so‘zlarini tadqiqot kontekstida juda tez-tez uchraydigan so‘zlar deb hisoblash mumkin, ular mavzularni aniqlashda qo‘shimcha qiymat bermaydi. Tasodifiy shovqin bo‘lishi mumkin bo‘lgan tez-tez uchraydigan so‘zlarni olib tashlash tavsiya etiladi.

(e) kamdan-kam hollarda. Hujjatda kamdan-kam uchraydigan atama bo‘lishiga yo‘l qo‘ymaslik uchun so‘zning minimal sonini belgilashingiz kerak bo‘lishi mumkin. Mavzuni bir necha yuzlab hujjatlarni modellashtirishda, agar natijalar shovqinli bo‘lsa yoki tahlil juda uzoq davom esa, kamida ikkita hujjatdan boshlash va ularni mos ravishda moslashtirish foydali bo‘ladi. Namuna minglab hujjatlardan iborat bo‘lgan hollarda, beshta hujjat kabi yuqori minimumdan boshlash yaxshidir. Ko‘pgina hollarda shovqinni kamaytirish va hisoblash tezligini oshirish uchun kamida beshta hujjat qo‘llaniladi.

(f) Stemming lemmatizasiya. Tahlilni osonlashtirish uchun stemming yoki lemmatizasiya qilish kerak. Masalan, “Imkoniyat” va “imkoniyat”lar “imkoniyat” bilan almashtiriladi. Ushbu bosqichdan oldin bir nechta dastlabki tahlillarni o‘tkazish tavsiya etiladi. Stemming keyingi tahlil uchun tavsiya etiladi, ammo tadqiqotchilar stemming chiqarilgan xulosalarga ta’sir qiladimi yoki yo‘qligini tekshirishlari kerak.

(g) Uni-/bi-/trigramlar. Bigramlardan foydalanish odatda so‘zlar sumkasini yengib o‘tishga yordam berib, tahlil qilishda yordam beradi. Misol uchun, faqat "SHimoliy" va "Karolina" o‘rniga "SHimoliy Karolina" dan foydalanishingiz mumkin. Biroq, uni, bi yoki trigramlardan foydalanganda topilmalar o‘zgarishini aniqlash uchun bu taxminni har bir tahlilda sinab ko‘rish mumkin.

4-qadam: Mavzuni modellashtirish

a) tez-tez ishlatiladigan so‘zlar. Matndagi "muhim" va tez-tez uchraydigan so‘zlar takrorlanuvchi jarayon orqali aniqlanishi kerak, bunda to‘xtash so‘zlarini qo‘shish va olib tashlash va bunday qadamlar paydo bo‘ladigan so‘z turlarini o‘zgartirish yoki o‘zgartirishni baholash uchun boshqa dastlabki ishlov berish bosqichlari.

(b) Mavzular soni. 1 dan 100 tagacha mavzuni ko‘rib chiqish tavsiya etiladi; paydo bo‘ladigan mavzular soni oxir-oqibatda mavzularning talqin qilinishi va parsimonlik zarurati bilan ta’sir qilishi kerak. Bir qator mavzular bo‘yicha tadqiqotlar iterativ tarzda amalga oshirilishi kerak.

(v) tarmoq strukturasini o‘rganish. Mavzular tarmog‘i muayyan mavzular qanchalik bog‘liqligini ko‘rish imkonini beradi. Bu paydo bo‘lgan konstruksiyalarning o‘lchovliligini baholash uchun foydalidir. Mavzu tarmog‘ini xaritalashda minimal korrelyasiya 0,10 dan boshlash tavsiya etiladi. Biroq, chegarani, ayniqsa mavzular soniga qarab sozlash kerak bo‘ladi (ko‘proq mavzular pastroq qiymatni talab qiladi va aksincha).

(d) ish ta’rifi. Mavjud adabiyotlardan konstruktiv ta’riflarni aniqlash yoki so‘zlar ro‘yxati uchun so‘zlarni tanlashni osonlashtirish uchun ishchi ta’rifni ishlab chiqish kerak.

Matndagi xatolarni tuzatish algoritmlari

Tabiiy tilni qayta ishlash (NLP) algoritmlari inson tilidagi ma’lumotlarni, shu jumladan tuzilmagan matn ma’lumotlarini qayta ishlash uchun ishlatiladi. Bugungi kunda NLP algoritmlari til qoidalari, statistik yondashuvlar va sun’iy intellekt yondashuvlari asosida ishlab chiqilgan. Til qoidalariga asoslangan yondashuv asosida NLP vazifalari va til korpusidagi tasniflash operasiyalarining lingvistik asoslarini shakllantirish amalga oshiriladi. Statistik algoritmlar mashinalarga inson tillarini o‘qish, tushunish va ma’no chiqarish imkonini beradi va katta hajmdagi matnlarni (katta ma’lumotlarni) qayta ishlashga asoslangan. Statistik algoritmlar nutqni aniqlash, mashina tarjimasi, hissiyotlarni tahlil qilish, tasniflash va matnni qazib olish kabi ko‘plab NLP vazifalarida qo‘llaniladi. Bugungi kunda CNN va RNN texnologiyalariga asoslangan mashinani o‘rganish (ML) algoritmlarini chuqur o‘rganish modellari mavjud NLP tizimlarini o‘rganishi va katta hajmdagi tuzilmagan matnni aniqroq qayta ishlashni ta’minlashi mumkin.

Bugungi kunda ma’lumotlarni qazib olish tabiiy tilni qayta ishlashning mashhur vazifalaridan biridir. Matnning semantikasini (mazmunini) tushunish, uni matn sifatida ifodalash va qayta ishlash uchun tasniflash, qismlarga ajratish kabi harakatlar bajarilishi kerak.

Kompyuterlar va odamlarga tabiiy til yordamida ma’lumot almashish imkonini beruvchi mexanizmlarni ishlab chiqadilar. Kompyuterlar NLP tufayli inson tilini o‘qishi, sharhlashi, tushunishi va javob berishi mumkin, ishlov berish odatda mashinaning aql darajasiga bog‘liq. NLP algoritmlari odamning xabarlarini ular uchun mazmunli raqamli ma’lumotlarga aylantiradi. Jarayonlar va ularning algoritmlari qanday ishlashini tushunish muhimdir. Zamon bilan hamnafas bo‘lish, bu texnologiyalar imkoniyatlaridan unumli foydalanish zarur.

NLPda hal qilinayotgan muammolar orasida eng muhimlari quyidagilardir (1-rasm) [10],[11]:

- NLP texnologiyalarini ishlab chiquvchilar uchun belgilangan birinchi klassik vazifa;

- grammatika va imloni tekshirish (grammatika va afsun checking) – o‘zbekcha matnlarning imlosini avtomatik tekshirish va tuzatish;
- matn tasnifi (matn klassifikatsiyasi) – matn semantikasini aniqlash;
- ma’lum ob’ektlarni identifikasiya qilish (nomli ob’ekt recognition, NER) - muhim ob’ektlarni aniqlash;
- umumlashtirish - matnni soddalashtirilgan shaklda umumlashtirish;
- matn yaratish (matn avlod) AI tizimlarini yaratishda foydalaniladigan vazifalardan biridir;
- mavzuni modellashtirish (mavzuni modellashtirish) - katta matnlardan yashirin mavzularni ajratib olish.

2-rasm. NLP vazifalarining murakkabligi[12]

Shuni ta’kidlash kerakki, barcha zamonaviy va tegishli NLP vazifalari ko‘pincha interaktiv AI tizimlarini yaratishga to‘g‘ri keladi: chatbotlar. Bu inson so‘rovlarini dasturiy ta’minot bilan birlashtirishga yordam beradigan muhit (tizim).

Bundan tashqari, mobil qurilmalarda matn xatolarini tuzatish jarayonlari paydo bo‘ladi. Bunday holda, foydalanuvchi kursorni xato joyiga ko‘chirish orqali xatoni tuzatishi mumkin. Bunday jarayonlar ko‘p matnli so‘zlarga sarflanadigan vaqtni ko‘paytirishga olib keladi. Tekshiruv jarayonini yaxshilashga yordam beradigan uchta yangi matnni tuzatish usuli ko‘rib chiqiladi: Drag-n-Drop, Drag-n-Throw va Magic Key.

- **Drag-n-Drop**. Bu usul foydalanuvchiga kursorni xato ustiga olib borish va uni tuzatish imkonini beradi.
- **Drag-n-Throw.** Ushbu usul foydalanuvchiga xato matnining dastlabki ro‘yxatini klaviatura so‘rovlari ro‘yxatidan olish imkonini beradi.
- **Sehrli kalit.** Ushbu usul foydalanuvchiga tuzatish kiritish imkonini beradi va neyron tarmoq tomonidan aniqlangan xatolar ro‘yxatini ko‘rsatadi.

Mobil qurilma foydalanuvchilarga yozgan matnni xato holatiga qaytarish uchun Backspace tugmachasini surish imkonini berdi va xatoni tuzatgandan so‘ng,

Backspace tugmasini yana bosish orqali ushbu matnni tiklang. Xato joylarini aniqlash uchun ushbu usul lug‘atdagi matn va so‘zlarni tahrirlash masofasini solishtirish uchun eugene Mers algoritmidan foydalanishi mumkin. Xatolarni aniqlash algoritmi bu ishning asosiy cheklovidir: u faqat imlo xatolarini aniqlaydi. U grammatik xatolar yoki so‘zlarning noto‘g‘ri ishlatilishini aniqlamasligi mumkin.

Xatolarni tuzatish uchun NLP algoritmlari

Drag-n-Throw va Magic Key qo‘llaniladigan tuzatishlar asosida yuzaga kelishi mumkin bo‘lgan xatolarni topish uchun tabiiy tilni qayta ishlash (NLP) neyron tarmoqlaridan foydalanadi. SHuning uchun biz tegishli usullar haqida qisqacha ma’lumot beramiz. An’anaviy xatolarni tuzatish algoritmlari tuzatish bo‘yicha tavsiyalar berish uchun N-gramm masi va qatorni tahrirlash masofasidan foydalanadi. Asl satrdagi har bir so‘z uchun avvalo lug‘atdan mosliklarni qidiradi va har bir mumkin bo‘lgan moslikni N-grammli ma’lumotlar to‘plamidagi chastotasi va qatorni moslashtirish algoritmi asosida almashtiradi. Eng yuqori matn tuzatish sifatida tavsiya etiladi. So‘nggi paytlarda chuqur neyron tarmoqlari NLP tadqiqotlarida ularning umumlashtirilishi va an’anaviy algoritmlarga qaraganda ancha yaxshi ishlashi tufayli mashhurlikka erishdi. Konvolyusion neyron tarmoqlari (CNN) va takroriy neyron tarmoqlari (RNN) NLP vazifalari uchun keng qo‘llaniladi va ko‘pincha kodlovchi-dekoder deb ataladigan tuzilishga amal qiladi, bu yerda modelning bir qismi kirish matnini xususiyat vektoriga kodlaydi va keyin vektorni dekodlaydi.

Foydalanilgan adabiyotlar:

1. Mamatov, N. S., Turg‘unova, N. M., Turg‘unov, B. X., & Samijonov, B. N. (2025). Algorithm for analysis of texts in Uzbek language. In Artificial Intelligence and Information Technologies (pp. 511-515). CRC Press.
2. Cassia Valentini-Botinhao, Xin Vang, Shinji Takaki, Junichi Yamagishi. Speech enhancement for a noise-tolerant text-to-speech synthesis system using deep recurrent neural networks [Onlayn]
3. Agiomyrgiannakis, Rob Klark, Rif A. Saurous. Tacotron: towards end-to-end speech synthesis [Onlayn].
4. Vey Ping, Kainan Peng, Endryu Gibianskiy, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonatan Rayman, Jon Miller. Deep Voice 3: Convolutional Sequence Learning (Onlayn) with text-to-speech conversion.
5. Jonatan Shen, Ruoming Pang, Ron J. Vayss, Mayk Shuster, Navdip Jaytli, Zongheng Yang, Zhifeng Chen, Yu Chjan, Yuxuan Vang, RJ Skerri-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yongxui Vu. Natural TTS Synthesis by WaveNet Conditioning on Mel Spectrogram Projections [Onlayn].

6. Quinn, K. M., Monro, B. L., Colaresi, M., Crespin, M. H. va Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. American Journal of Politology, 54, 209–228.

7. Niyozmatova, N. A., Mamatov, N. S., Dusonov, X. T., Samijonov, B. N., & Samijonov, A. N. MFCC-GMM Method for Speaker Identification by Voice.

8. Baumer, E. P., Mimno, D., Guha, S., Quan, E. va Gay, G. K. (2017). Comparing Grounded Theory and Subject Modeling: Extreme Divergence or Unlikely Convergence? Journal of the Association for Information Science and Technology, 68, 1397–1410.

9. Strauss, A. va Corbin, J. (1998). Foundations of qualitative research: Techniques and procedures for developing grounded theory (2nd ed.). A Thousand Oaks: Sage.

10. Reynard, J. C. (2008). Introduction to Communication Studies (4th ed.). Boston: McGraw-Hill.

11. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys (CSUR), 34, 1–47.

12. Zilola Xusainova Yuldashevna. NLPning zamonaviy algoritmlari va konsepsiyalari // журнал искусство слова №1 (2023) https://doi.org/10.5281/7679911