MATNNI TASNIFLASH VA ULARNING SOHALAR KESIMIDA QO‘LLANILISH MASALASI

Xusainova Zilola Yuldashevna,
Filologiya fanlari bo‘yicha falsafa doktori
xusainovazilola@navoiy-uni.uz
ToshDO‘TAU

Shirinboyeva Marjona,
Filologiya yo‘nalishi talabasi
shirinboyevamarjona36@gmail.com
ToshDO‘TAU

Annotatsiya. Matnni tasniflash – bu matnlarni mazmun, uslub, maqsad yoki tarkibiy jihatdan guruhlash jarayoni. Zamonaviy axborot texnologiyalari rivojlanishi natijasida matnni tasniflash usullari ko‘plab sohalarda qo‘llanilmoqda. Xususan, hujjatlarni avtomatik ajratish, elektron pochta spamlarini aniqlash va ijtimoiy tarmoqlardagi kontent monitoringi kabi vazifalarda keng foydalaniladi. Tibbiyotda, yuridik sohada matn tasnifi muhim ahamiyat kasb etadi. Matnni tasniflashda asosan sun’iy intellekt va mashinali o‘qitish texnologiyalari muhim yo‘nalishlaridan biri hisoblanadi. Bu jarayon matnlarni avtomatik ravishda oldindan belgilangan toifalarga ajratishni anglatadi. Masalan, ijtimoiy tarmoqlardagi postlarni sentimentini aniqlashda, ya’ni ijobiy, salbiy yoki neytral deb tasniflashda qo‘llaniladi. Shu bilan birga, NLP metodlari matnlarning ma’nosini chuqur tahlil qilish imkonini beradi. Har bir sohada matn tasniflashning o‘ziga xos xususiyatlari va talablariga to‘g‘ri tanlangan algoritm ishlab chiqiladi.

Abstract. Text classification is the process of grouping texts by content, style, purpose, or structure. As a result of the development of modern information technologies, text classification methods are used in many areas. In particular, they are widely used in tasks such as automatic document sorting, detecting email spam, and monitoring content on social networks. Text classification is of great importance in medicine and the legal field. Text classification is mainly one of the important areas of artificial intelligence and machine learning technologies in text classification. This process involves automatically dividing texts into predefined categories. For example, it is used to determine the sentiment of posts on social networks, that is, classify them as positive, negative, or neutral. At the same time, NLP methods allow for a deep analysis of the meaning of texts. In each area, an algorithm is developed that is correctly selected for the specific characteristics and requirements of text classification.

Аннотация. Классификация текстов — это процесс группировки текстов по содержанию, стилю, назначению или структуре. В результате развития современных информационных технологий методы классификации

текстов нашли применение во многих областях. В частности, он широко используется в таких задачах, как автоматическая сортировка документов, обнаружение спама в электронной почте и мониторинг контента социальных сетей. Классификация текстов важна в медицине и юридической сфере. Классификация текстов — одна из важнейших областей применения технологий искусственного интеллекта и машинного обучения. Этот процесс подразумевает автоматическую классификацию текстов по предопределенным категориям. Например, он используется для определения тональности постов в социальных сетях, то есть для их классификации как положительных, отрицательных или нейтральных. В то же время методы НЛП позволяют проводить глубокий анализ смысла текстов. В каждой области разрабатывается алгоритм, адаптированный под конкретные характеристики и требования классификации текста.

Kalit so‘zlar. *Matnni tasniflash, NLP, ML, axborotni tartiblash, sentiment tahlil, sohaviy tasnif.*

Matnni tasniflash – bu matnlarni ma’lum belgilarga ko‘ra guruhlarga ajratish va ularni avtomatik tarzda toifalash jarayonidir. Bugungi kunda axborot texnologiyalari jadal rivojlanishi natijasida matn ko‘rinishidagi ma’lumotlar hajmi keskin oshmoqda. Ushbu ma’lumotlar oqimini matnlarni tizimli ravishda tasniflash va ularni sohaga oid kategoriyalarga ajratish dolzarb ilmiy-texnik masalaga aylanmoqda. Matnni tasniflash turli mezonlar (mazmun, struktura, til xususiyatlari) asosida matnlarni sistematik ravishda guruhlash jarayonidir. Ushbu jarayon tabiiy tilni qayta ishlash, mashinali o‘qitish va sun’iy intellekt texnologiyalarining rivojlanishi bilan yanada takomillashib bormoqda. Biroq, sifatli ma’lumotlarning yetishmasligi va til xilma-xilligi bu sohada muhim muammolar sifatida qolmoqda. Ushbu maqolada matnni tasniflash jarayonining nazariy asoslarini va ularning turli sohalar kesimidagi amaliy qo‘llanilishini ilmiy asosda tahlil qilishga qaratilgan.

Matnni tasniflash zamonaviy axborot texnologiyalari va sun’iy intellekt sohasining muhim tadqiqot yo‘nalishlaridan biri sifatida shakllangan. Ushbu jarayon matnli ma’lumotlarni oldindan belgilangan toifalar yoki kategoriyalarga avtomatik ravishda ajratishni nazarda tutadi[1]. Matnni tasniflash tabiiy tilni qayta ishlash (Natural Language Processing) va mashinali o‘qitish (Machine Learning) texnologiyalariga tayangan holda amalga oshiriladi. Har bir sohada matnlarning semantik va kontekstual xususiyatlarini hisobga olgan holda moslashtirilgan modellardan foydalaniladi[2]. Shuningdek, matnni tasniflash texnologiyalari axborot oqimini boshqarish, tezkor qaror qabul qilish va umumiy samaradorlikni oshirishda muhim ahamiyat kasb etmoqda.

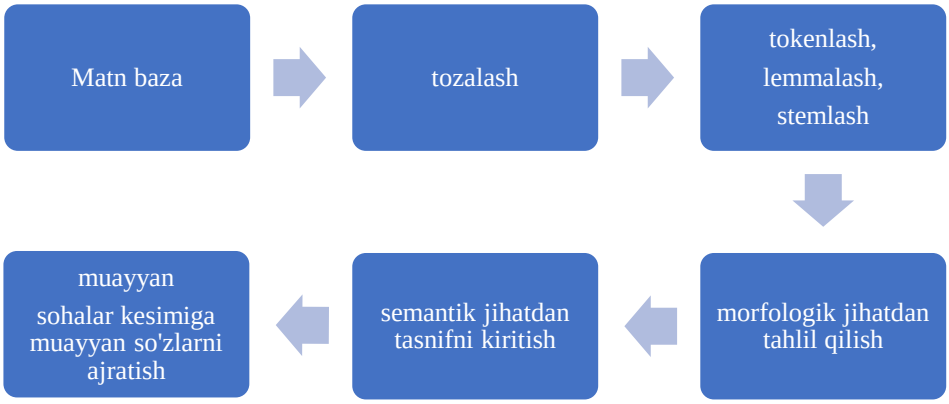
Ayniqsa, matnlarning sohalarda tasniflashda ancha vaqt talab etadi, ammo uning hozirda aniq yechimi mavjud emas. Bo‘lsa ham tahlillarning samaradorlik

ko‘rsatkichi 100% natija bermaydi. Bu kabi muammolar bor ekan ularni tahlil qilish va yechim topishga bo‘lgan qiziqlash va talablar ortib boradi. Matnlarni tasniflash jarayonida ularni avtomatik ravishda toifalarga ajratish uchun ma’lum so‘zlarni matn tarkibidagi kalit so‘zlarga qarab uning sentiment scorini ijobiy, salbiy kabi guruhlarga yoki qaysi sohaga oid ekanligini aniqlab berishda muhim ahamiyatga ega. Masalan: internet tarmog‘ida jurnalistik matnlar, ilmiy yoki badiiy matnlar soniya sayin ko‘payib bormoqda ularni ajratish ko‘proq vaqt talab etadi. Jurnalistik matnlarni ajratishning o‘zi ham bir necha turlarga ajratiladi. Misol tariqasida sport, musiqa, tibbiyot, adabiyot, tabiat haqidagi matnlar. Bular tarkibida ma’lun so‘zlarni ayni shu soha ekanligini o‘rganib ularni ajratadi. Misol uchun, “Bugun yurtimizning janubi-sharqiy qismida yomg‘irli ob-havo bo‘lishi kutilmoqda” gapida ob-havo, yomg‘ir va janubi-sharqiy so‘zlari orqali bu matnning ob-havo haqida ma’lumot berayotganini bilishimiz mumkin.

Matnni tasniflash bo‘yicha olib borilgan dastlabki tadqiqotlar ko‘proq qoidaga asoslangan (rule-based) yondashuvlarga tayangan[3]. Ushbu usullar matndagi muhim kalit so‘zlar va sintaktik tuzilmalarga asoslanib tasniflashni amalga oshirgan. Biroq, qoidaga asoslangan yondashuvlar katta hajmdagi va murakkab ma’lumotlar bilan ishlashda o‘zining cheklovlarini namoyon qilgan. Keyingi bosqichda statistik metodlar va mashinaviy o‘rganish asosidagi yondashuvlar rivojlandi[4]. Xususan, Naive Bayes, Support Vector Machines (SVM) va Decision Trees kabi modellar matnni tasniflash samaradorligini oshirdi. So‘nggi yillarda chuqur o‘rganish (Deep Learning) usullari, xususan, neyron tarmoqlar va transformer modellar (masalan, BERT) matnlarni yanada aniqroq tasniflash imkonini berdi. Masalan, Devlin va boshqalar[5] tomonidan ishlab chiqilgan BERT (Bidirectional Encoder Representations from Transformers) modeli kabi pre-train qilingan til modellarining rivojlanishi matnni kontekstual ravishda chuqur tahlil qilishga imkon yaratdi. Matnni tasniflash bo‘yicha tadqiqotlar tibbiyot, moliya va ijtimoiy tarmoqlar monitoringi kabi ko‘plab sohalarda qo‘llanila boshlandi. Masalan, klinik yozuvlardan kasallikni aniqlash, kreditni baholash kabi amaliy holatlarni kuzatishimiz mumkin. Biroq mavjud metodologiyalarning cheklovlari ham mavjud. Dastlabki usullar kontekst va semantika chuqur tahlil qilinmagani uchun murakkab matn tuzilmalari bilan ishlashda samarasiz bo‘lgan. Shuningdek, kam resursli tillar uchun matnni tasniflash modellari hali ham yetarlicha rivojlanmagan. Mavjud nazariyalar asosan katta hajmdagi strukturalangan ma’lumotlarga tayanadi, bu esa kichik korpuslar uchun qiyinchilik tug‘diradi. Mashinali o‘qitsh modellarining samaradorligi esa ko‘pincha katta hajmdagi, yaxshi belgilangan ma’lumot to‘plamlariga bog‘liqdir. Deep learning asosidagi modellar esa hisoblash resurslariga va ma’lumotga bo‘lgan yuqori talab tufayli cheklovlarga ega. Shuningdek, soha kesimida matnni tasniflashda maxsus domenga moslashtirilgan modellarni ishlab chiqish zarurati tug‘iladi.

Ko‘plab tadqiqotlar umumiy tasniflash modellariga qaratilgan, ma’lum bir soha uchun moslashgan tasniflagichlar hali to‘liq mukammallashmagan. Shuningdek, matnlarning noaniqligi, teglanmagan va imlo xatolariga nisbatan modellar samaradorligini oshirish masalasi yetarlicha yechilmagan.

Sohalar bo‘yicha matn tasniflashning bizga afzal tomonlari shundaki, bu jarayon tez va unumdor bo‘lishi hamda ishchi kuchiga bo‘lgan talabning kamayishidir.



Tadqiqotlardan olingan natijalar matnni tasniflash texnologiyalari turli sohalarda axborot boshqaruvini samarali tashkil etishga xizmat qilishini ko‘rsatadi. Shuningdek, har bir soha uchun moslashtirilgan model va algoritmlarni tanlash samaradorlikni oshiradi. Jurnalistika, onlayn harid, sport, tibiiyot, tilshunoslik va matematika sohalarida aniqlik talablari yuqori bo‘lganligi sababli, chuqur o‘rganishga asoslangan (Deep Learning) yondashuvlar ustunlik qilmoqda.

Online savdo sohasida

Agar biz biror narsani online xarid qilmoqchi bo‘lsak mahsulot bizga qay holatda kelishi, bizga mahsulot qaday ta’sir qilishi, ayniqsa, kosmetika mahsulotlari, kiyim-kechak, qurilish mollari, hozirgi paytda keng ommalashgan audio kitoblar va doimiy reklamalarda aylanayotgan turli xil dori vositalarini sifatini aniqlashdagi mijozlarning fikrlarini bilishda yordam beradi. Bu savdo sohasidagi mijozlardan biri agar “Bu mahsulot menga yoqdi”, “Yaxshi”, “Mahsulot menga manzur keldi”, “Zo‘r”, “Mahsulot sifatli va arzon” kabi fikrlar bildirsa demak bu fikrlarni biz ongli ravishda iliq fikrlar ekanligini anglaymiz. Ammo, buni sun’iy intellektga avtomatik ajratish uchun avvalo, o‘rgatish kerak. “Yoqmadi”, “Menga uncha manzur emas”, yana salbiy baho ifodalovchi bir qancha stikerlardan foydalanish orqali salbiy munosabatlarni bilib olishimiz mumkin. Bu fikrlar munosabatni bildirsa, yana bir jihati borki “buyurtma qilmoqchiman”, “yetkazib berildi”, “mahsulot haqida quyidagi havoladan bilib olishingiz mumkin” kabi bir qancha avtomatik javob berish mumkin bo‘ladi. Keling ijtimoiy tarmoqlardagi ba’zi savdo kanallaridagi commentlardan misol tariqasida ko‘rib chiqamiz.

1. Kompyuter sotadigan Instagram commentida “**qmatku**” tarzidagi fikr qoldirilgan ushbu fikrda biz imloviy xatoni ko‘rishimiz mumkin ammo bu kabi holatlar yana topiladi va buni bazaga kiritish orqali salbiy ma’no berishini bilishimiz mumkin. Yana bir qancha misollar keltiramiz:

“**Qimmat** va **bir tiyingayam qimmat** bu keca 3 millionga coria 5 abrativka 8 xotira 512 360 graduagaca iciladi va sensor 12 avlod.”, — salbiy.

“Men ishlataman **Juda yaxshi** laptop”, — ijobiy.

2. Kiyim do‘kon kommentidan:

🔥🔥 — ijobiy

😔😔 — salbiy

“**juda yaxshi** narx ham sifati ham yaxshi”, ijobiy.

“Yetkazib berish **uncha yaxshi emas** ekan 1 hafta deganda **illa** oldim”, ushbu fikrda salbiy baho bilan birgalikda sheva so‘z **“illa”** ham ishtirok etgan.

Jurnalistika sohasida

Matnlarni mavzu ko‘lami va ma’no jihatdan salbiy yoki ijobiy guruhlarga tasniflashda kerak bo‘ladi. Masalan; “Keldiyorovaga bu mukofot Parij-2024 Olimpiadasida yaponiyalik taniqli dzyudochi Uta Abe ustidan qozonilgan g‘alaba uchun taqdim etildi”, kabi mavzuga doir maqolani biz **“dyuzdochi”, “Uta Abe”** kabi teglar orqali sport mavzuda ekanligini, “Domla sharhlarni o‘qibdi-da, osmonga qarab charaqlagan quyoshni Alisherga, qora bulutni esa Husayn Mirzoga ko‘rsatib, Alisherga “Ofarin” debdi”, jumlasidan adabiyotga oid bo‘lgan “Alisher”, “Husayn Mirzo” teglari orqali bilishimiz, “O‘zbekistonda hafta davomida iliq va asosan quruq ob-havo bo‘lishi kutilmoqda.” jumlasidagi “iliq”, “ob-havo” va “quruq” teglari orqali ob-havo mavzusida ekanligini va albatta, bu kabi keltirilgan fikrlar ijobiy ekanligini aniqlashga yordam bersa, “2025-yil 24-fevral. Bugun Rossiyaning Ukraina zaminiga boshlagan keng ko‘lamli urushiga 3 yil to‘ldi. Bu voqelik nafaqat ikki malakatda balki, butun dunyo bo‘ylab sado berdi: minglab insonlar halok bo‘ldi, ko‘z o‘ngimizda shahar vayronaga aylandi, boshqa davlatlar o‘z tomonlarini hamon aniqlamoqda.”, jumlasidan esa biz bu mavzudagi **“urush”** tokeni orqali urush mavzusida ekanligini va “halok bo‘ldi”, “vayronaga aylandi” kabi so‘zlar orqali esa salbiy ma’no ifodalayotganini bilishimiz mumkin.

Tibbiyot sohasida

Kimdir sog‘lig‘idan shikoyat qilib,

– haroratning ko‘tarilishi, bosh og‘rig‘i, isitma, yo‘tal, tomoq og‘rig‘i kabi belgilar borligini aytsa, bu orqali bemorda grippga chalinganligini osongina aniqlash va unga kerakli dori hamda muolajalar haqida ma’lumot berish mumkin bo‘ladi.

– holsizlik, tez-tez kasal bo‘lish, tezda vazn yo‘qotish, sariqlik, yo‘tal va ovozning bo‘g‘ilishi kabi belgilar esa saratondan darak berishi hamda zudlik bilan davolash ishlarini boshlash haqida ma’lumot va tavsiyalar beradi. Bu kabi misollardan ularni “tibbiy terapiya” “nevrologiya” kabi bolimlarga xos ekanligini bilishingiz mumkin.

– yana bir tomondan “qabulga yozildi”, “x- navbatdagi bemor”, “qon tahlili” kabi tokenlar orqali bir qancha zaruriy ma’lumotlarga ega bo‘lish mumkin.

Tilshunoslik sohasida

Lug‘atlar, imlo tekshiruvi, bosh harflar imlosi, chiziqcha va tire bilan bog‘liq bilan muammolar, punktuatsion belgilar, transkripsiya kabi bir qancha masalalar bor.

– Lug‘atchilikning eng ommalashgan turi ingliz-o‘zbek, rus-o‘zbek, turk-o‘zbek kabi tarjima lug‘atlari hisoblanadi. Ammo arab, fors va tojik kabi izofali tillarda bu biroz oqsayotganini ko‘rishimiz mumkin. Izofalar bizni tilimizga xos bo‘lmaganligi sababli tarjima sohasida ularda qator muammolar kelib chiqyapti.

– Eski o‘zbek tilidan yurtimizda ancha yillar foydalanilganini hisobga olsak, bu yozuvdagi asarlarni hozirgi o‘zbek adabiy tiliga o‘girish va bu jarayondagi ishlar hali o‘z yechimiga to‘la ega emasligini ko‘rimiz mumkin. To‘g‘ri eski o‘zbek tili va yozuvi haqida bir qancha qo‘llanma, lug‘atlar ishlab chiqigan bo‘lsa hamki, ushbu masalani avtomatik hal qilish kerak ekanligi yana bir masaladir.

– Punktuatsion belgilarni to‘g‘ri ishlatilishida hamma ham xatolarga yo‘l qo‘yishi mumkin. Ammo tez-tez ishlatilinuvchi nuqta, vergul, ikki nuqtani ishlatishda ularni uyushiq holatda kelganda verguldan foydalanish, kesimlikni shakklantiruvchi qo‘shimchalardan so‘ng nuqtaning ishlatilishi va uyushiq so‘zlardan oldin,ko‘chirma gapdan oldin qo‘llangan muallif gapidan so‘ng ishlatishni hal etish masalasi ham bor.

– Tire va chiziqcha bilan bog‘liq muammo ham muhim masalalardan biridir. Qay holatda tire, qay holatda chiziqchaning ishlatilishiga ko‘pchilik ikkilanib qoladi. Tire punktuatsion, chiziqcha esa orfografik belgidir. Juft va takror so‘zlar orasida chiziqcha ishlatilishi kabi qator muammolar o‘z yechimini kutmoqda.

Algebra-matematika sohasida

Matn tasnifi berilgan misol, masala, tenglama, tengsizlik, funksiya kabi tur va oson, qiyin, o‘rtacha darajalarga ajratishda muhim ahamiyatga egadir.

Dialektologiya sohasida

Sheva matnlaridagi so‘zlarni adabiy tilga o‘girish, unli harflarni old va orqa qatorga, undoshlarni esa qay holatda talaffuz etilgan bo‘lsa uning adabiy tildagi muqobiliga almashtirishda kerak bo‘ladi.

XULOSA

Matn tasniflash turli sohalarda ma’lumotlarni tahlil qilish va qaror qabul qilish jarayonlarini optimallashtirishning muhim vositasi bo‘lib, AI texnologiyalarining rivojlanishi bilan uning ahamiyati tobora ortib bormoqda. Matnlarni tasniflash kelgusida ham sun’iy intellekt yutuqlari bilan uyg‘un holda rivojlanib, axborot jamiyatining samaradorligini oshirishda muhim rol o‘ynashi kutilmoqda.

Foydalanilgan adabiyotlar:

1. Aggarwal, C. C. (2012). Mining Text Data. Springer.

2. Camacho-Collados, J., & Pilehvar, M. T. (2018). From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. Journal of Artificial Intelligence Research.

3. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys.

4. Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. ECML.

5. Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.

6. B. Elov, Sh. Hamroyeva, R. Alayev, Z. Xusainova, U. Yodgorov – O‘zbek tili korpusi matnlarini qayta ishlash usullari – Raqamli Transformatsiya va Sun’iy Intellekt ilmiy jurnali, October 2023 – Volume 1, Issue 3.

7. Xusainova, Z., & Yangibayeva, S. (2024). Til korpusi turlari. Uzbekistan: Language and Culture, 2(2).

8. Elov, B. B., Hamroyeva, S., & Xusainova, Z. (2022). NLP (tabiiy tilga ishlov berish) ning vazifalari va zamonaviy yondashuvlar. TerDU, filologik tadqiqotlar: til, adabiyot, ta’lim, 5-6.